Original Research

Self-Other Agreement in Multisource Feedback: The Influence of Doctor and Rater Group Characteristics

Martin J. Roberts, MSc; John L. Campbell, MD; Suzanne H. Richards, PhD, MBPsS; Christine Wright, PhD

Introduction: *Multisource feedback (MSF) ratings provided by patients and colleagues are often poorly correlated with doctors' self-assessments. Doctors' reactions to feedback depend on its agreement with their own perceptions, but factors influencing self-other agreement in doctors' MSF ratings have received little attention. We aimed to identify the characteristics of doctors and their rater groups that affect self-other agreement in MSF ratings.*

Methods: *We invited 2454 doctors to obtain patient and colleague feedback using the UK General Medical Council's MSF questionnaires and to self-assess on core items from both patient (PQ) and colleague (CQ) questionnaires. Correlations and differences between doctor, patient and colleague mean feedback scores were examined. Regression analyses identified the characteristics of doctors and their rater groups that influenced self-other score agreement.*

Results: *1065 (43%) doctors returned at least one questionnaire, of whom 773 (73%) provided self and patient PQ scores and 1026 (96%) provided self and colleague CQ scores. Most doctors rated themselves less favourably than they were rated by either their patients or their colleagues. This tendency to underrate performance in comparison to external feedback was influenced by the doctor's place of training, clinical specialty, ethnicity and the profile of his/her patient and colleague rater samples but, in contrast to studies undertaken in nonmedical settings, was unaffected by age or gender.*

Discussion: *Self-other agreement in MSF ratings is influenced by characteristics of both raters and ratees. Managers, appraisers, and others responsible for interpreting and reviewing feedback results with the doctor need to be aware of these influences.*

Key Words: *multisource feedback, patient surveys, peer assessment, self-assessment, self-other agreement, continuous professional development*

Introduction

Multisource feedback is an established method of assessing workplace performance and its suitability as a tool for assessing the professional performance of doctors has been widely studied.[1,2] While the potential of such feedback to improve doctors' performance has been recognized, the extent of its impact on practice is influenced by a range of factors related to the content of the feedback and the context and format in which it is delivered and received.[3–7] One such factor is "self-other agreement": the congruence between external assessments of performance and an individual's self-assessment (in the sense described by Eva and Regehr as "an unguided, personal, summative assessment of one's own level of ability or performance").[8] While numerous studies have reported on the incongruity between doctors' perceptions of their own performance and external measures,[8–10] realistic self-assessment, conceptualized more broadly as a process of collecting, interpreting, assimilating, and responding to information from multiple sources on one's own performance, has been regarded as a cornerstone of self-directed continuing professional development.[5,11–13] Self-other agreement is an important part of this self-assessment process: studies of feedback given to doctors have reported that disagreement with negative feedback can not only cause distress to the

Disclosures: JLC is an advisor to the GMC and has recieved only direct costs associated with presentation of this work. The other authors have no conflicts of interest.

Mr. Roberts: Department of Primary Care, Peninsula College of Medicine and Dentistry, University of Exeter; *Dr. Campbell:* Department of Primary Care, Peninsula College of Medicine and Dentistry, University of Exeter; *Dr. Richards:* Department of Primary Care, Peninsula College of Medicine and Dentistry, University of Exeter; *Dr. Wright:* Department of Primary Care, Peninsula College of Medicine and Dentistry, University of Exeter.

Correspondence: Martin J. Roberts, University of Exeter, St Lukes Campus, Magdalen Road, Exeter EX1 2LU, United Kingdom; e-mail: martin.roberts@pms.ac.uk.

• Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/chp.21162

doctors involved but may adversely affect the likelihood that they will act upon it.[13–17]

Research in nonmedical settings has found self-other agreement in feedback ratings to be associated with a range of demographic, professional, personality, and cultural characteristics of ratees and their rater groups.[18–23] Moreover, good agreement has been found to be positively related to aspects of job performance such as effectiveness and leadership ability.[21,24] A number of studies have examined the association of demographic and other factors with the accuracy of medical students' self-assessments when compared to "gold standards" such as faculty ratings or objective performance measures.[9,25] As far as we are aware, however, no detailed investigation of the association of such factors with self-other agreement in multisource feedback ratings of doctors has been undertaken. In light of the growing use of multisource feedback in revalidation and other processes requiring assessment of doctors' professional performance, the present study aims to address this gap in the literature.

We analyzed data from a recent large-scale study of multisource feedback in fully trained doctors.[26] Having identified the demographic and professional characteristics of the assessed doctors and their rater groups that influenced variation in patient and colleague feedback ratings,[27] our aim in the present study is to identify the characteristics that are associated with variation in self-other agreement. In examining such variation, it is important not only to identify group differences but to trace these differences back to their origin in the separate ratings.[20] For example, if a particular group of doctors were found to overrate their patient consultation skills, it would be important to ascertain whether the difference was caused by those doctors rating themselves more favorably than other doctors or by patients rating them less favorably (or both). Our analysis therefore aims to elucidate the characteristics that significantly influence variation in self-other agreement.

Methods

An evaluation of the use of multisource feedback questionnaires developed for the UK General Medical Council (GMC) provided patient, colleague, and self-assessed ratings on the professional performance of fully trained doctors. The two primary feedback instruments were a patient questionnaire (PQ) containing 9 core performance evaluation items and a colleague questionnaire (CQ) containing 18 such items. All items were rated using 5-point scales. A self-assessment questionnaire incorporating both patient-related and colleague-related items was sent to participating doctors. The doctor questionnaire also included demographic (age, gender, ethnicity, place of medical qualification) and professional context (specialty, contractual role, time in role, locum status, intensity of patient contact) items. Full details of the patient and colleague instruments and their psychometric properties have been published elsewhere.[26,28]

Self-Other Agreement Scores

Four measures of performance were derived for each doctor. From the patient questionnaires we derived a "patient-PQ score" for each doctor provided that, in line with our original instructions to participants, at least 22 patient questionnaires had been returned. The patient-PQ score was derived by first calculating a mean rating for each core item where at least 6 patients had returned a valid rating and then calculating the mean of these item means where more than half were available. A parallel approach was adopted to derive a "colleague-CQ score" where at least 8 colleague questionnaires had been returned and more than half of the possible 18 core item means were available. The self-assessment questionnaire furnished a mean score for the 9 patient-related items (self-PQ score) and a mean score for the 18 colleague-related items (self-CQ score). These scores were only derived where more than half of the relevant items were scored. We calculated the reliability (Cronbach's alpha) of these 4 performance scores.

We then calculated, where possible, two measures of self-other agreement for each doctor: a PQ agreement score, defined as their self-PQ score minus their patient-PQ score, and a CQ-agreement score equal to their self-CQ score minus their colleague-CQ score.

Relative to assessments provided by their patients or colleagues, doctors could thus have under-rated their own performance, resulting in negative PQ or CQ agreement scores. Alternatively, they may have relatively overrated performance, resulting in positive PQ or CQ agreement scores. Under- or overrating could thus derive from variations in either self-assessment or the assessments provided by others or a combination of both. We considered the concept of under- or overrating by doctors as reflecting their self-assessment relative to the assessments provided by their patient or colleague raters, rather than as a comparison with a predefined "gold standard."

Data Analysis

Statistical analysis was conducted in PASW Statistics 18. Due to skewness and influential outliers in the patient- and colleague-derived scores, we calculated Spearman's rank correlation coefficients between the self-PQ, patient-PQ, self-CQ, and colleague-CQ scores. Paired sample t-tests were used to test for differences between the 2 patient-related scores, and between the 2 colleague-related scores. We used multiple linear regression analysis to determine which demographic, professional context, and rater sample variables were independently associated with the PQ and CQ

DOI: 10.1002/chp

agreement scores and with the underlying self, patient, and colleague scores. Potential predictors of (independent variables associated with) the PQ agreement score, the self-PQ score, and the patient-PQ score (the dependent variables) were entered into all 3 multiple regression models if a bivariate regression of any one of these 3 scores on the predictor resulted in a p-value below 0.10. Potential predictors of the three corresponding colleague-related scores were selected in the same way. In interpreting the multiple regression analyses, we regarded variables as significant independent predictors of the outcome variable if, after correcting for other variables in the model, the resulting p-value was less than 0.05.

To explain these independent predictors further we disaggregated the source of the PQ or CQ agreement score, determining whether variation in agreement was attributable to differences in self-assessment, differences in patient- or colleague-derived assessments, or both.

The methodology of the evaluation study from which this analysis is derived was considered by the Devon and Torbay NHS Research Ethics Committee but judged not to require a formal ethics submission.

Results

A total of 2454 doctors from 11 UK trust settings were invited to participate in the study; 1065 (43% participation rate) returned at least 1 questionnaire and, of these, 773 (73%) provided self-PQ and patient-PQ scores (Cronbach's alpha = 0.90 and 0.87, respectively) and 1026 (96%) provided self-CQ and colleague-CQ scores (Cronbach's alpha = 0.91 and 0.94, respectively).

The doctors' two self-ratings were strongly correlated (Spearman's rho = 0.815, $p < 0.001$), indicating a possible halo effect in these scores, but neither was correlated with the patient or colleague ratings (rho = −0.013 to 0.069). The patient ratings were positively, though weakly, correlated with the colleague ratings (rho = 0.320, $p < 0.001$). The majority of doctors underrated themselves when compared to the assessments of their patients and colleagues (paired sample t-tests, $p < 0.0001$ in both cases) and consequently 82% of PQ and 86% of CQ agreement scores were negative (PQ agreement score mean = −0.47, standard deviation = 0.46, N = 773, skewness = −0.331; CQ agreement score mean = −0.47, standard deviation = 0.44, N = 1026, skewness = −0.087).

Agreement With Patient Ratings

Variables not included in the linear regression model for the PQ agreement score were: the doctor's frequency of contact with patients, the doctor's length of time in their current contractual role, and the proportion of patients returning their questionnaire by post. Variables removed from the model due to multicollinearity were the proportion of patients under 15 years of age, the proportion in Asian ethnic groups, and the proportion whose questionnaire was completed by a proxy. The final model (TABLE 1) accounted for 12.4% of the variation in PQ agreement scores. Independent predictors of those scores were the doctor's region of primary medical qualification and the proportion of young (under 21years) and old (60 years or over) patients.

Given the overall tendency of doctors to rate themselves less favorably than did their patients, it was evident (FIGURE 1) that, relative to their patients' assessments, UK-trained doctors underrated themselves more severely than doctors trained elsewhere. Doctors trained outside the United Kingdom tended to either rate themselves more highly than their UK-trained peers (non-UK European-trained doctors) or to receive less favorable patient feedback scores (South Asian–trained doctors) or both (doctors trained in "other" regions), resulting in higher PQ agreement scores (TABLE 1). The proportion of younger (under 21 years) and older (60 years or over) patients in the rater sample both had a negative effect on doctors' PQ agreement scores. This was due to lower self-assessments by those doctors who had higher proportions of patient respondents in these age groups.

Agreement With Colleague Ratings

Variables not included in the linear regression model for the CQ agreement score were the doctor's frequency of contact with patients, the doctor's length of time in his or her current contractual role, and the percentage of colleague raters who were trainee doctors. Variables removed from the model due to multicollinearity were the proportion of colleagues in Asian ethnic groups, the proportion in daily contact with the doctor, the proportion in administration/managerial roles, and the proportion in allied health care roles. The final model (TABLE 2) explained 15.6% of the variation in CQ agreement scores. Independent predictors of the CQ agreement score were the ethnicity, region of primary medical qualification and specialty group of the doctor and the proportion of their colleague sample who were qualified doctors or who reported more frequent professional contact with the doctor.

Compared to UK- and South Asian–trained doctors, doctors qualifying from other regions tended to underrate themselves less severely relative to the assessments provided by their colleagues. This was a result of both higher self-assessments and lower colleague feedback scores (TABLE 2). Doctors from Asian ethnic groups tended to receive less favorable colleague feedback and consequently underrated themselves less severely relative to the assessments provided by their colleagues than those from White ethnic groups (FIGURE 2).

TABLE 1. Effect of Doctor and Patient Sample Characteristics on Agreement, Self- and Patient-Assessed Scores for Patient-Related Items

	Subgroup	PQ agreement			Self-PQ score[a]			Patient-PQ score[a]		
	size[b]	P[c]	B[d]	(95% CI)[e]	P[c]	B[d]	(95% CI)[e]	P[c]	B[d]	(95% CI)[e]
Doctor characteristics										
Gender		0.502								
Male	466		Ref							
Female	248		0.027	(−0.051, 0 .105)						
Age group		0.121								
20–39	132		Ref							
40–49	319		0.066	(−0.027, 0.158)						
50–59	198		0.069	(−0.034, 0.171)						
60 and over	65		0.173	(0.031, 0.315)						
Ethnic group		0.436								
White	568		Ref							
Asian	109		0.072	(−0.091, 0.235)						
Other	37		0.088	(−0.075, 0.250)						
Region of PMQ[f]		0.000			0.000			0.000		
United Kingdom	536		Ref			Ref			Ref	
EEA[g] (non-UK)	37		0.296	(0.146, 0.446)		0.265	(0.116, 0.414)		−0.031	(−0.068, 0.006)
South Asia	77		0.247	(0.051, 0.442)		0.164	(−0.029, 0.358)		−0.082	(−0.130, −0.035)
Other	64		0.269	(0.136, 0.402)		0.212	(0.080, 0.343)		−0.057	(−0.090, −0.024)
Clinical specialty group		0.903								
General practice	331		Ref							
Medical	199		−0.018	(−0.120, 0.083)						
Surgical	124		0.028	(−0.079, 0.134)						
Psychiatry	24		0.054	(−0.139, 0.246)						
Other	36		0.010	(−0.152, 0.172)						
Locum status		0.611								
Nonlocum	692		Ref							
Locum	22		0.051	(−0.147, 0.250)						
Contractual role		0.605								
Consultant / GP	619		Ref							
Other	95		0.028	(−0.077, 0.132)						
Patient sample characteristics[h]										
% of patients who are female		0.503	−0.007	(−0.028, 0.014)						
% of patients who are under 21		0.045	−0.026	(−0.051, −0.001)	0.038	−0.027	(−0.052, −0.002)	0.824	−0.001	(−0.007, 0.006)
% of patients who are over 60		0.005	−0.027	(−0.046, −0.008)	0.009	−0.025	(−0.044, −0.006)	0.350	0.002	(−0.002, 0.007)
% of patients whose ethnic group is 'white'.		0.589	−0.010	(−0.046, 0.026)						
% of patients whose visit is 'very important'		0.989	0.000	(−0.031, 0.031)						
% of patients who are seeing their usual doctor		0.638	−0.004	(−0.018, 0.011)						

[a]Regression models for these variables included all predictor variables but we report results only for significant predictors of the agreement score. [b]Sample size = 714 doctors. [c]*P*-value for significance of predictor. [d]Regression coefficient (Ref denotes the reference category). [e]Wald-based confidence interval. [f]Primary Medical Qualification. [g]European Economic Area. [h]Coefficients expressed as the increase in the outcome variable per 10% increase in the predictor.

DOI: 10.1002/chp

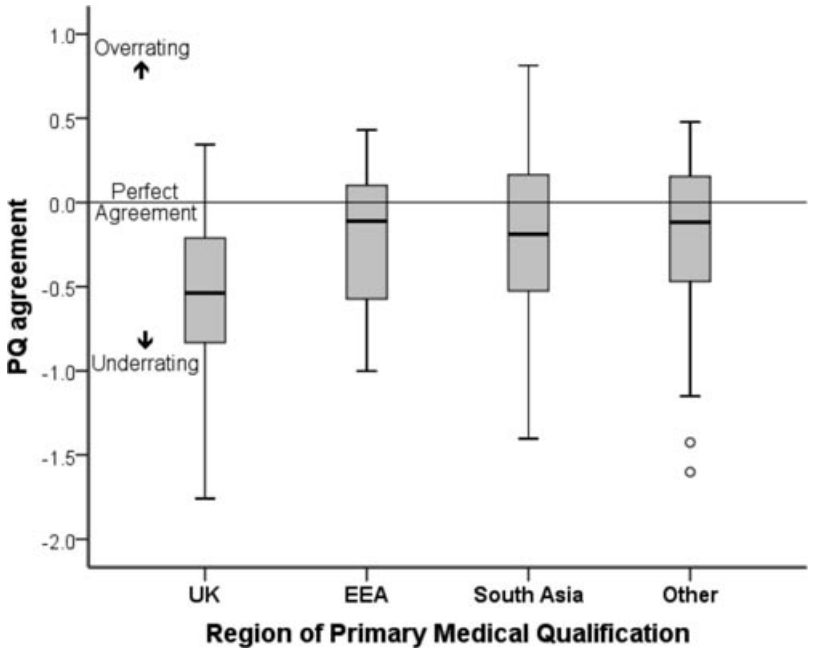

FIGURE 1. Boxplots of Patient Agreement Score by Doctor's Region of Primary Medical Qualification (° indicates an outlier in the data, greater than 1.5 times the interquartile range below the lower quartile. N = 714.)

In contrast, hospital doctors in the "medical" specialty group received more favorable colleague feedback and hence underrated themselves more severely relative to the assessments provided by their colleagues than did general practitioners, surgeons and psychiatrists (TABLE 2).

The profile of the colleague sample also influenced self-other agreement. Doctors with a higher proportion of colleagues who reported more frequent professional contact with the doctor received more favorable colleague feedback scores and hence underrated themselves more severely. Doctors with a greater proportion of medically qualified colleagues in their rater sample tended to self-assess themselves more highly, resulting in greater congruence with their colleague ratings.

Discussion

Despite the growing use of multisource feedback as a tool to aid the professional development of doctors and the recognized effect of self-other agreement on reactions to such feedback, the factors that may influence this agreement (or lack of it) have hitherto received little attention. Based on a large sample of fully trained doctors, our study has examined the correlations and differences among self, patient, and colleague ratings of doctors' professional performance. We have identified demographic and professional characteristics of the doctors and their rater groups that influenced self-other agreement in feedback scores.

In common with studies of self-assessment amongst medical students and of feedback on the performance of practicing doctors,[10,29–31] we found doctors' self-ratings to have no correlation with those of other rater groups, though this finding is not universal.[9] Patient ratings were positively correlated with colleague ratings, though, like those reported elsewhere,[31,32] these correlations were weak. We cannot say, however, that one group of raters has accurately assessed the doctor and that the other two sources are "wrong": the notion that any rater group provides a gold standard against which to judge the accuracy of others is misleading.[29] The patient-, colleague- and self-ratings are simply differing sources that contribute to the data-gathering stage of the doctor's self-assessment process while the extent of self-other agreement may influence the interpretation, assimilation, and response stages. Our study therefore supports the view that a particular strength of multisource feedback is to provide distinct perspectives on the assessed doctor that combine to give "a more complete picture of performance."[1]

We found that self-other agreement was influenced by the doctor's region of primary medical training and, in the case of doctor-colleague agreement, by their ethnic background. These observations suggest the existence of a cultural component to self-other agreement in the medical profession. We report these findings for the first time in relation to the professional practice of doctors, although similar observations have been made in nonmedical settings.[19–22] We also found variation in agreement scores to be associated with aspects of the

DOI: 10.1002/chp

TABLE 2. Effect of Doctor and Colleague Sample Characteristics on Agreement, Self- and Colleague-Assessed Scores for Colleague-Related Items

	Subgroup size[b]	CQ Agreement			Self-CQ score[a]			Colleague-CQ score[a]		
		P[c]	B[d]	(95% CI)[e]	P[c]	B[d]	(95% CI)[e]	P[c]	B[d]	(95% CI)[e]
Doctor characteristics										
Gender		0.767								
Male	619		Ref							
Female	330		0.009	(−0.052, 0.071)						
Age group		0.078								
20–39	182		Ref							
40–49	429		0.022	(−0.050, 0.095)						
50–59	260		−0.015	(−0.097, 0.067)						
60 and over	78		0.130	(0.006, 0.254)						
Ethnic group		0.049			0.316			0.043		
White	750		Ref			Ref			Ref	
Asian	146		0.152	(0.015, 0.290)		0.076	(−0.057, 0.208)		−0.077	(−0.137, −0.016)
Other	53		0.097	(−0.032, 0.225)		0.075	(−0.049, 0.200)		−0.021	(−0.078, 0.036)
Region of PMQ[f]		0.000			0.002			0.001		
United Kingdom	707		Ref			Ref			Ref	
EEA[g] (non-UK)	49		0.258	(0.138, 0.378)		0.169	(0.053, 0.285)		−0.089	(−0.142, −0.036)
South Asia	107		0.124	(−0.036, 0.283)		0.087	(−0.067, 0.241)		−0.036	(−0.106, 0.034)
Other	86		0.205	(0.100, 0.311)		0.146	(0.044, 0.248)		−0.060	(−0.106, −0.013)
Clinical specialty group		0.008			0.141			0.000		
General practice	355		Ref			Ref			Ref	
Medical	320		−0.077	(−0.152, −0.002)		0.007	(−0.065, 0.079)		0.084	(0.051, 0.117)
Surgical	169		0.038	(−0.047, 0.123)		0.097	(0.014, 0.179)		0.059	(0.021, 0.096)
Psychiatry	53		0.099	(−0.030, 0.229)		0.049	(−0.076, 0.174)		−0.050	(−0.107, 0.007)
Other	52		−0.022	(−0.147, 0.103)		0.037	(−0.083, 0.158)		0.059	(0.004, 0.114)
Locum status		0.914								
Nonlocum	926		Ref							
Locum	23		0.010	(−0.164, 0.184)						
Contractual role		0.309								
Consultant/GP	825		Ref							
Other	124		0.043	(−0.040, 0.127)						
Colleague sample characteristics[h]										
% of colleagues who are under 30 years old		0.670	−0.010	(−0.053, 0.034)						
% of colleagues who are 60 or more years old		0.233	−0.020	(−0.054, 0.013)						
% of colleagues who are female		0.375	0.010	(−0.012, 0.032)						
% of colleagues whose ethnic group is 'white'		0.068	−0.023	(−0.048, 0.002)						
% of colleagues who are doctors (inc. trainees)		0.032	0.027	(0.002, 0.052)	0.074	0.022	(−0.002, 0.046)	0.338	−0.005	(−0.016, 0.006)
% of colleagues who currently work with the doctor		0.127	−0.019	(−0.043, 0.005)						
% of colleagues who are/were in daily or weekly contact with the doctor		0.024	−0.022	(−0.041, −0.003)	0.467	−0.007	(−0.025, 0.012)	0.000	0.015	(0.007, 0.023)
% of colleagues who returned a paper version of the questionnaire		0.342	−0.006	(−0.018, 0.006)						

[a]Regression models for these variables included all predictor variables but we report results only for significant predictors of the agreement score. [b]Sample size = 949 doctors. [c]*P*-value for significance of predictor. [d]Regression coefficient (Ref denotes the reference category). [e]Wald-based confidence interval. [f]Primary Medical Qualification. [g]European Economic Area. [h]Coefficients expressed as the increase in the outcome variable per 10% increase in the predictor.

DOI: 10.1002/chp

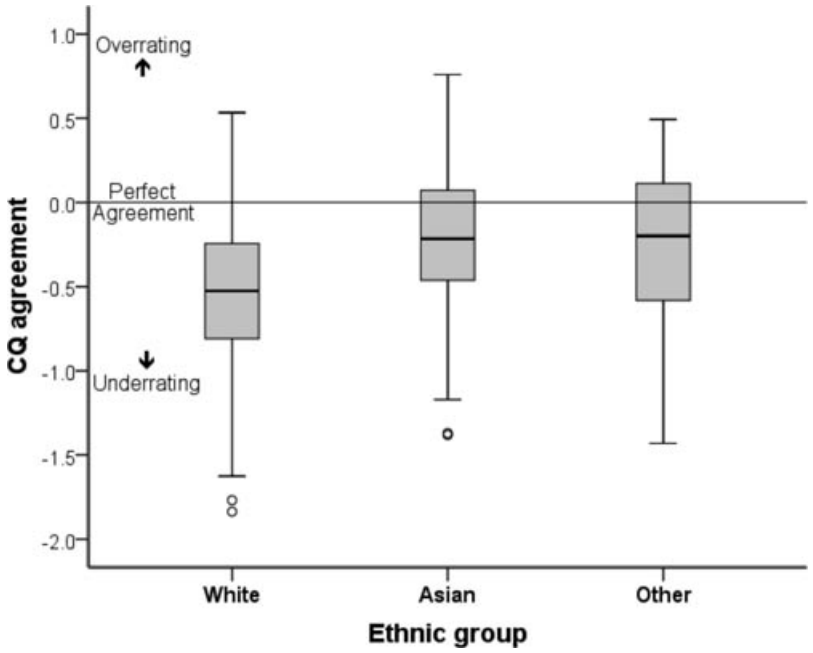

FIGURE 2. Boxplots of Colleague Agreement Score by Doctor's Ethnic Group (° indicates an outlier in the data, greater than 1.5 times the interquartile range below the lower quartile. N = 949.)

doctor's working environment (clinical specialty, profile of the patient and colleague respondents), adding support to the view that self-assessment is a complex process. Our finding of variation in self-colleague agreement across different clinical specialties contrasts with the results of Violato and colleagues, who found no differences in agreement among psychiatrists, pediatricians, and internal medicine specialists.[33] Our results give no support however, to the notion that self-other agreement might improve as doctors gain professional experience: neither age nor seniority (as indicated by "contractual role") were significantly associated with variation in agreement scores. Gender was also not a significant predictor of this variation. These results for age and gender differ from those found in nonmedical settings, where males and older ratees have consistently been found more likely to overrate their own performance in comparison to feedback ratings provided by others.[20–22]

Negative reactions to multisource feedback are more likely to occur when external ratings are less favorable than anticipated.[14,15,17,34] However, like several previous studies,[30,32,35,36] we found that most doctors underrated themselves in comparison to their patient and colleague ratings. This suggests that only a small minority of doctors will react negatively to such feedback. Our results indicate, however, that certain subgroups of doctors (eg, those doctors who trained outside of the United Kingdom, those with a greater proportion of medically qualified colleagues in their rater sample) are likely to be overrepresented in that minority. Doctors who overrate their professional performance relative to the assessments provided by others may represent potential risks to patient safety through a relative lack of insight into their lower performance as assessed by external sources such as patients or colleagues. It is, however, more difficult to interpret the potential implications of serious underrating, which was much more common in our data. Whether serious underrating reflects a worrying lack of insight on the part of some doctors or whether they are simply "gaming," either to avoid appearing overoptimistic about their own performance or to circumvent disappointment, is a matter of speculation.

The study has a number of limitations. In reality, multisource feedback is not simply reported as a single score averaged across all questionnaire items and other features of the way in which the feedback might be presented (eg, mean scores for individual items, allocation into norm-referenced percentile bands, comparison with specialty benchmarks, inclusion of anonymized free text comments) may influence practitioners' responses. The present study casts no direct light on these alternative forms of feedback, though it is reasonable to infer from our findings that self-other agreement in those forms could also be influenced by characteristics of the doctors and their rater groups. The particular characteristics that influenced self-other agreement in our UK-based study may not, however, apply elsewhere in the world. A further limitation of the study is the volunteer nature of the sample, which, albeit participating and nonparticipating doctors were demographically similar,[26] may have led to

DOI: 10.1002/chp

underrepresentation of poorly performing doctors. This may, in part, explain the predominance of underrating in our study as poorer performers are recognized as being more likely to overrate their own performance.[9,37,38] Our choice of the arithmetic difference between the self- and other assessment scores as our measure of self-other agreement may also have influenced our findings. This choice was based on simplicity and precedent but is not the only possibility: other measures such as absolute differences and correlations have been used elsewhere.[24,39] Violato and colleagues, for example, used differences in mean percentile ranks to examine doctor-peer rating agreement.[33] In common with many UK-developed multisource feedback instruments, the GMC colleague questionnaire was designed for use by both peers (other doctors) and coworkers (other health care professionals, managers, and administrators). This approach can be justified by the argument that splitting colleagues into peers and coworkers is an oversimplification of the complex web of professional relationships that surround the assessed doctor. Rather than develop two separate questionnaires, a wider perspective is gained by offering all items to all colleagues, allowing them to rate whichever aspects of the doctor's professional performance that they feel able to rate. While this approach is partly justified by our previous finding that the professional makeup of a doctor's colleague group does not affect their mean rating,[27] it does have the disadvantage of making comparisons with studies that use separate peer and coworker instruments less straightforward.

Leadership has been identified as important for quality improvement in health care and for the ongoing development of the medical profession in the 21st century, leading to calls for greater emphasis on the importance of medical leadership and to initiatives to improve leadership and management skills by embedding them in medical education curricula.[40–42] We referred in our introduction to the finding, in nonmedical settings, that close agreement with others' feedback ratings is positively correlated with leadership ability.[21,24] This is not a connection that the current study was able to investigate, but it may, in light of current interest in medical leadership, provide a fruitful avenue for further research into self-other agreement. The proportions of variance in self-other agreement scores explained by our regression models may also indicate scope for further research. These proportions were relatively low (12.4% and 15.6% for the patient and colleague models, respectively), suggesting the existence of factors associated with variation in self-other agreement that were not included in our models.

Conclusions

Self-assessment is widely seen as an important process in the largely self-regulating profession of medicine, yet doctors are recognized to be poor self-assessors.[8–11] Regular validation of self-assessments by comparison with external assessments such as those arising from multisource feedback has therefore been recommended as a vital part of the ongoing cycle of continuing professional development.[12,33,43,44] Eva and Regehr emphasized the importance of "understanding factors that influence our ability to absorb these external sources of feedback in developing a coherent self-awareness of our strengths and weaknesses."[8] We have shown that discrepancies between patient and colleague feedback ratings and doctors' self-ratings, which are undoubtedly a factor influencing the ability to absorb feedback, may be influenced by characteristics of both the doctors and of their rater groups. In situations where, as recommended by Sargeant and others,[5,13,14,45] doctors are assisted in assimilating their feedback by the provision of "facilitated reflection on feedback," both providers and facilitators must be aware of these potential influences. This has important implications for the growing number of countries that have incorporated, or are intending to incorporate, multisource feedback in their processes for revalidation or recertification of practicing doctors.

Lessons for Practice

- Doctors' self-ratings were both uncorrelated with and less favorable than the ratings provided by their patients or colleagues.
- The tendency to underrate performance in comparison to external feedback was stronger in UK-trained doctors than in those trained abroad and was influenced by the doctor's ethnicity, clinical specialty, and the profile of his/her patient and colleague rater samples.
- Self-other agreement, which can influence reactions to feedback, may therefore be affected by cultural background and clinical setting.
- In contrast to studies conducted in nonmedical settings, the tendency to underrate was unrelated to the doctor's age or gender.

Acknowledgments

The authors wish to thank all doctors, their patients, and colleagues who contributed to this study and the senior management teams at hosting organizations who supported the work. CFEP-UK organized and managed the recruitment and data collection aspects of the survey. The study was funded by

DOI: 10.1002/chp

the UK General Medical Council as an unrestricted research award.

References

1. Lockyer J. Multisource feedback in the assessment of physician competencies *J Contin Educ Health Prof*. 2003;23(1):4–12.
2. Dubinsky I, Jennings K, Greengarten M, Brans A. 360-degree physician performance assessment. *Healthc Q*. 2010;13(2):71–76.
3. Veloski J, Boex JR, Grasberger MJ, Evans A, Wolfson DB. Systematic review of the literature on assessment, feedback and physicians' clinical performance: BEME Guide No. 7. *Med Teach*. 2006;28(2):117–128.
4. Eva KW, Armson H, Holmboe ES, et al. Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. *Adv Health Sci Educ*. 2012;17:15–26.
5. Archer JC. State of the science in health professional education: effective feedback. *Med Educ*. 2010;44(1):101–108.
6. Castanelli D, Kitto S. Perceptions, attitudes, and beliefs of staff anaesthetists related to multi-source feedback used for their performance appraisal. *Br J Anaesth*. 2011;107(3):372–377.
7. Sargeant JM, Macleod T, Sinclair DE, Power M. How do physicians assess their family physician colleagues' performance? Creating a rubric to inform assessment and feedback. *J Contin Educ Health Prof*. 2011;31(2):87–94.
8. Eva KW, Regehr G. "I'll never play professional football" and other fallacies of self-assessment. *J Contin Educ Health Prof*. 2008;28(1):14–19.
9. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence. *JAMA*. 2006;296(9):1094–1102.
10. Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. *Acad Med*. 1991;66(12):762–769.
11. Eva KW, Regehr G. Self-Assessment in the health professions: a reformulation and research agenda. *Acad Med*. 2005;80(10):S46–S54.
12. Galbraith RM, Hawkins RE, Holmboe ES. Making self-assessment more effective. *J Contin Educ Health Prof*. 2008;28(1):20–24.
13. Sargeant JM, Armson H, Chesluk B, et al. The processes and dimensions of informed self-assessment: A conceptual model. *Acad Med*. 2010;85(7):1212–1220.
14. Sargeant J, Mann K, Sinclair D, Van der Vleuten CPM, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv Health Sci Educ*. 2008;13(3):275–288.
15. Lockyer JM, Violato C, Fidler H. Likelihood of change: a study assessing surgeon use of multisource feedback data. *Teach Learn Med*. 2003;15(3):168–174.
16. Sargeant JM, Mann K, van der Vleuten CPM, Metsemakers J. "Directed" self-assessment: practice and feedback within a social context. *J Contin Educ Health Prof*. 2008;28(1):47–54.
17. Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: perceptions of credibility and usefulness. *Med Educ*. 2005;39(5):497–504.
18. Smither JW, London M, Richmond KR. The relationship between leaders' personality and their reactions to and use of multi-source feedback: a longitudinal study. *Group Organ Manage*. 2005;30(2):181–210.
19. Atwater LE, Wang M, Smither JW, Fleenor JW. Are cultural characteristics associated with the relationship between self and others' ratings of leadership? *J Appl Psychol*. 2009;94(4):876–886.
20. Ostroff C, Atwater LE, Feinberg BJ. Understanding self-other agreement: a look at rater and ratee characteristics, context, and outcomes. *Pers Psychol*. 2004;57(2):333–375.
21. Fleenor JW, Smither JW, Atwater LE, Braddy PW, Sturm RE. Self-other rating agreement in leadership: a review. *Leadership Q*. 2010;21(6):1005–1034.
22. Brutus S, Fleenor JW, McCauley CD. Demographic and personality predictors of congruence in multi-source ratings. *J Manage Develop*. 1999;18(5):417–435.
23. Heidemeier H, Moser K. Self-other agreement in job performance ratings: a meta-analytic test of a process model. *J Appl Psychol*. 2009;94(2):353–370.
24. Atwater LE, Ostroff C, Yammarino FJ, Fleenor JW. Self-other agreement: does it really matter? *Pers Psychol*. 1998;51(3):577–598.
25. Colthart I, Bagnall G, Evans A, et al. The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME Guide no. 10. *Med Teach*. 2008;30(2):124–145.
26. Wright C, Richards SH, Hill JJ, et al. Multisource feedback in evaluating the performance of doctors: the example of the UK General Medical Council Patient and Colleague Questionnaires. *Acad Med*. 2012;87(12):1668–1678.
27. Campbell JL, Roberts MJ, Wright C, et al. Factors associated with variability in the assessment of UK doctors' professionalism: analysis of survey results. *BMJ*. 2011;343:d6212.
28. Campbell JL, Hill JJ, Hobart J, et al. GMC Multi-Source Feedback Study. Scientific report of the Main Survey (2008–10): Executive Summary. General Medical Council, London, United Kingdom, 2012. http://www.ncbi.nlm.nih.gov/pubmed/22331475 . Accessed August 17, 2012.
29. Ward M, Gruppen L, Regehr G. Measuring self-assessment: Current state of the art. *Adv Health Sci Educ*. 2002;7(1):63–80.
30. Kenny DA, Veldhuijzen W, Weijden T, et al. Interpersonal perception in the context of doctor-patient relationships: a dyadic analysis of doctor-patient communication. *Soc Sci Med*. 2010;70(5):763–768.
31. Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol RP, Lombarts KM. Evaluation of physicians' professional performance: an iterative development and validation study of multisource feedback instruments. *BMC Health Serv Res*. 2012;12:80.
32. Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. *Acad Med*. 2002;77(10):S64–S66.
33. Violato C, Lockyer J. Self and peer assessment of pediatricians, psychiatrists and medicine specialists: implications for self-directed learning. *Adv Health Sci Educ*. 2006;11(3):235–244.
34. Brett JF, Atwater LE. 360° feedback: accuracy, reactions, and perceptions of usefulness. *J Appl Psychol*. 2001;86(5):930–942.
35. Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. *BMJ*. 2003;326(7388):546–548.
36. Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: the Physician Achievement Review. *Can Med Assoc J*. 1999;161(1):52–57.
37. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol*. 1999;77(6):1121–1134.
38. Hodges B, Regehr G, Martin D. Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it. *Acad Med*. 2001;76(10)(Supplement):S87–S89.
39. Fitzgerald JT, White CB, Gruppen LD. A longitudinal study of self-assessment accuracy. *Med Educ*. 2003;37(7):645–649.
40. Tooke J. Aspiring to excellence: findings and recommendations of the Independent Inquiry into Modernising Medical Careers. MMC Inquiry, London, United Kingdom, 2008. http://www.mmcinquiry.org.uk/Final_8_Jan_08_MMC_all.pdf. Accessed Aug 18, 2012.

41. Clark J, Armit K. Attainment of competency in management and leadership: no longer an optional extra for doctors. *Clin Governance Int J*. 2008;13(1):35–42.
42. NHS Institute for Innovation and Improvement and Academy of Medical Royal Colleges. *Medical Leadership Competency Framework. Enhancing Engagement in Medical Leadership*. 3rd ed. NHS Institute for Innovation and Improvement, Coventry, United Kingdom, 2010. Available at: http://www.institute.nhs.uk/images/documents/Medical%20Leadership%20Competency%20Framework%203rd%20ed.pdf. Accessed August 18, 2012.
43. Sargeant JM, Mann KV, van der Vleuten CPM, Metsemakers J. "Directed" self-assessment: practice and feedback within a social context. *J Contin Educ Health Prof*. 2008;28(1):47–54.
44. Duffy FD, Lynn LA, Didura H, et al. Self-assessment of practice performance: development of the ABIM Practice Improvement Module (PIMSM). *J Contin Educ Health Prof*. 2008;28(1):38–46.
45. Overeem K, Wollersheim H, Driessen E, et al. Doctors' perceptions of why 360-degree feedback does (not) work: a qualitative study. *Med Educ*. 2009;43(9):874–882.

DOI: 10.1002/chp