Christine Wright, John Campbell, Luke McGowan, Martin J Roberts, Di Jelley and Arunangsu Chatterjee

Interpreting multisource feedback:

online study of consensus and variation among GP appraisers

Abstract

Background
GPs collect multisource feedback (MSF) about their professional practice and discuss it at appraisal. Appraisers use such information to identify concerns about a doctor's performance, and to guide the doctor's professional development plan (PDP).

Aim
To investigate whether GP appraisers detect variation in doctors' MSF results, and the degree of consensus in appraisers' interpretations of this information.

Design and setting
Online study of GP appraisers in north-east England.

Method
GP appraisers were invited to review eight anonymised doctors' MSF reports, which represented different patterns of scores on the UK General Medical Council's Patient and Colleague Questionnaires. Participants provided a structured assessment of each doctor's report, and recommended actions for their PDP. Appraiser ratings of each report were summarised descriptively. An 'agreement score' was calculated for each appraiser to determine whether their assessments were more lenient than those of other participants.

Results
At least one report was assessed by 101/146 appraisers (69%). The pattern of appraisers' ratings suggested that they could detect variation in GPs' MSF results, and recommend reasonable actions for the doctors' PDP. Increasing appraiser age was associated with more favourable interpretations of MSF results.

Conclusion
Although preliminary, the finding of broad consensus among GP appraisers in their assessment of MSF reports should be reassuring for GPs, appraisers, and employing organisations. However, if older appraisers are more lenient than younger appraisers in their interpretation of MSF and in the actions they suggest to their appraisees as a result, organisations need to consider what steps could be taken to address such differences.

Keywords
appraiser; general practitioners; multisource feedback; primary health care; revalidation.

INTRODUCTION

The UK General Medical Council (GMC) requires all practising doctors, over a 5-year cycle, to collect supporting information to demonstrate adherence to the principles described in *Good Medical Practice*.[1] Doctors are expected to reflect on this information and discuss it as part of their appraisal process.[2] The supporting information includes multisource feedback (MSF) on the doctor's practice obtained from colleagues and patients. MSF is viewed as a formative process, enabling individual doctors to identify where they may need to change their practice and to plan their future professional development.[2]

A number of questionnaires are available to support the collection of MSF. The GMC has developed its own Patient Questionnaire (PQ) and Colleague Questionnaire (CQ), which assess various aspects of professional practice.[3] When feedback has been collated, each doctor is provided with a personalised report, summarising (for each core PQ/CQ item): the distribution of ratings of the doctor's performance (5-point scales); a mean item percentage score; benchmark data derived from item percentage scores of other UK doctors; and the doctor's self-assessment rating. Free-text comments provided by the doctor, their patients, and colleagues are also presented.

There is evidence that the GMC questionnaires are acceptable for use within appraisal to provide formative feedback on a doctor's performance.[3] The resulting feedback can be complex, however, and should be interpreted with caution.[3,4] Benchmark data are predominantly derived from volunteer doctor samples, and are markedly skewed towards positive views of performance. Thus, an item score of 80–90% could still place a doctor in the lowest quartile when compared with their peers.[3] Furthermore, scores can be biased by factors associated with the individuals providing feedback or with the doctors themselves.[3,4]

Although the literature supports the use of MSF to improve practice,[5–8] a range of factors (relating to the individual doctor, their reaction to the feedback, and the availability of facilitation) may affect how a doctor uses the information to change their practice.[6–8]

GMC guidance recommends that doctors discuss their MSF with an individual trained in providing feedback (such as their appraiser). Appraisers are expected to make 'accurate and consistent judgements'[9] about supporting information

C Wright, PhD, research fellow; **J Campbell**, MD, FRCGP, professor, Primary Care Research Group, University of Exeter Medical School, Exeter. **A Chatterjee**, PhD, director; **L McGowan**, MSc, senior learning technologist, Technology Enhanced Learning for Medicine and Dentistry (TELMeD) Team, Plymouth University Peninsula Schools of Medicine and Dentistry and Peninsula College of Medicine and Dentistry, Plymouth. **MJ Roberts**, MSc, lecturer, Collaboration for the Advancement of Medical Education Research and Assessment (CAMERA), Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth. **D Jelley**, EdD, FRCGP, associate director, Health Education (North East and Cumbria), Newcastle.

Address for correspondence
John Campbell, Primary Care Research Group, University of Exeter Medical School, Smeall Building, St Luke's Campus, Magdalen Road, Exeter, Devon, EX1 2LU, UK.
E-mail: john.campbell@exeter.ac.uk

Submitted: 31 March 2015; **Editor's response:** 11 May 2015; **final acceptance:** 17 July 2015.

How this fits in

Doctors now collect and reflect on feedback from their patients and colleagues as part of the appraisal process. Little is known about how appraisers interpret multisource feedback (MSF) information. This study explored GP appraisers' interpretations of a purposively selected sample of MSF reports for eight doctors. The findings suggest that appraisers can detect variation in GPs' MSF results and suggest appropriate action based on these, but appraisers may vary in the leniency/stringency of their interpretation of MSF information.

to determine whether there are concerns about patient safety or the doctor's conduct or performance. Resources have been developed to support appraisers in the wider process of revalidation,[10,11] but these do not focus in detail on the interpretation of MSF. In one UK qualitative study,[5] appraisers of GPs reported difficulty in interpreting benchmark information; that is, whether PQ/CQ item scores falling in the lowest quartile benchmark band are indicative of GP performance that should give cause for concern.

Little is known about the consistency of interpretation of MSF by GPs and their appraisers. Research focusing on other 'high-stakes' performance-based assessments has observed examiner differences ('hawk–dove effect' or 'stringency/leniency effect')[12–15] that appear to be stable across time. In one UK study,[14] some examiners were observed to be more stringent (hawkish) in their assessment of candidates, and to require a higher level of performance for passing candidates than did other examiners. Although there was evidence that hawkishness correlated with examiner experience (number of candidates assessed) and ethnic origin, there was no evidence that it varied with examiner age or sex.[14,16] Other work in Canada suggests that individual examiners may be unaware of the extent of their stringency/hawkishness.[17]

Study aims

An online training resource was piloted to support the preparation of medical appraisers for their role in facilitating doctors' reflection on MSF, within the context of UK appraisal and revalidation. This study aimed to:

- assess participating appraisers' ability to detect variation in doctors' MSF scores;
- explore the degree of consensus between appraisers in their assessments of MSF results and actions they recommend; and
- examine variation between appraisers and identifying potential predictors of stringency in their interpretation of MSF.

METHOD

An online training resource was designed and constructed to provide GP appraisers with experience of interpreting MSF reports, and feedback on how their own interpretations compared with those of other appraisers. The design incorporated four clearly labelled sections: background information about the project; instructions on using the resources; access to eight MSF reports (labelled 'A' to 'H'); and a feedback function allowing appraisers to compare their own assessment of each MSF report with assessments submitted by other appraisers.

Each MSF report summarised feedback for one GP in the format described above (report data available from the author). Seven were real reports issued to UK GPs in earlier piloting of the GMC questionnaires.[3] At the end of that pilot work, standardised (Z) scores on the PQ and the CQ had been calculated for 402 doctors. A Z score below –1.96 was taken to indicate that the doctor's score fell in the lower tail of the distribution of doctor scores on the questionnaire (that is, their score was statistically outlying). Based on the doctor's PQ and CQ Z scores, their report was categorised into one of four groups:

(i) neither PQ nor CQ score statistically outlying;

(ii) PQ score statistically outlying but CQ score not statistically outlying;

(iii) CQ score statistically outlying but PQ score not statistically outlying; or

(iv) PQ and CQ scores both statistically outlying.

Reports available on the online training resource were purposively selected to represent different patterns of PQ and CQ scores (Table 1), were anonymised, and used with the doctors' explicit consent. Feedback indicative of poorer GP performance (group (iv) above) was rare in the earlier pilot study,[3] therefore report D was constructed to simulate such feedback. Appraisers who assessed the reports were unaware of the doctors' actual Z scores.

Appraisers were asked to review each MSF report and complete a 6-item online

Table 1. Overview of multisource feedback reports available for review by appraisers

Report	Outlying PQ score?[a]	Outlying CQ score?[a]	Item scores, *n*/*N* (%) PQ in lower quartile band[b]	PQ in upper quartile band[c]	CQ in lower quartile band[b]	CQ in upper quartile band[c]	Areas of concern highlighted in P or C free-text comments
A	No (Z = 0.29)	No (Z = 0.10)	1/9 (11)	3/9 (33)	3/18 (17)	3/18 (17)	None — all comments positive
B	**Yes (Z = –2.02)**	No (Z = –0.72)	8/9 (89)	0/9 (0)	4/18 (22)	3/18 (17)	None — all comments positive
C	No (Z = –1.41)	**Yes (Z = –2.45)**	7/9 (78)	0/9 (0)	18/18 (100)	0/18 (0)	Record keeping; prescribing (C)
D	**Yes (Z = –4.46)**	**Yes (Z = –2.19)**	5/9 (55)	0/9 (0)	11/18 (61)	0/18 (0)	Interpersonal skills (P,C)
E	No (Z = 0.12)	No (Z = 0.37)	0/9 (0)	0/9 (0)	1/18 (6)	6/18 (33)	None — all comments positive
F	No (Z = –1.63)	No (Z = –0.83)	9/9 (100)	0/9 (0)	9/18 (50)	3/18 (17)	Record keeping; aloofness (C)
G	No (Z = –1.08)	**Yes (Z = –1.93)**	6/9 (67)	0/9 (0)	13/18 (72)	0/18 (0)	Managing time/ workload (C)
H	No (Z = 0.69)	No (Z = 0.33)	0/9 (0)	7/9 (78)	1/18 (6)	3/18 (17)	None — all comments positive

[a]Outlying Patient Questionnaire (PQ) or Colleague Questionnaire (CQ) overall scores, which are emboldened in the table, were defined as those lying >1.96 standard deviations below the mean PQ or CQ overall score (standardised Z score < –1.96) calculated for all doctors who participated in GMC questionnaire pilot work.[3] [b]Number of PQ or CQ core items where the doctor's score fell in the lowest 25% of item scores achieved by doctors who participated in GMC questionnaire pilot work.[3] [c]Number of PQ or CQ core items where the doctor's score fell in the highest 25% of item scores achieved by doctors who participated in previous pilot work.[3] C = colleague. P = patient.

form. Three ordinal-scale items evaluated the appraisers' interpretation of the doctor's MSF report: an overall assessment of the report (5-point scale: 'Excellent' to 'Unsatisfactory'); their level of concern about the GP's performance (4-point scale: 'Not at all concerned' to 'Extremely concerned'); and the acceptability of the GP's performance (4-point scale: 'Clearly acceptable' to 'Clearly unacceptable') based on the content of their MSF report. Three categorical items indicated the actions appraisers would discuss during the GP's appraisal: repeating the patient/colleague surveys; specific actions/training for the doctor's personal development plan (PDP); and other possible actions. Responders could also add free-text comments about the MSF report or their recommended actions.

The process was repeated for each MSF report in turn and appraisers could choose the order in which they assessed reports. Assessments could be completed over a number of sessions but could not be amended once submitted. After submitting an assessment, appraisers could access the feedback function to view a summary of other appraisers' assessments of the same report.

Preliminary user-testing of the online training resource was conducted (July–October 2012) with three GP appraisers to check the acceptability of the registration process, training exercise, and supporting materials. Based on their feedback, changes were made to the training materials and web pages.

The revised training resource was made available to 235 GP appraisers from north-east England, in a series of waves (December 2012–November 2013). A panel of eight appraisers took part in the initial wave of recruitment and the panel's ratings of and comments about the constructed report (report D) suggested that this had face validity.

Appraisers were invited by the local appraisal lead to use the online resource as part of their continuing professional development. To register for an account, appraisers selected a username/password and provided brief demographic information. Accounts were individually verified and activated by the researcher, after which appraisers could access the eight MSF reports. Up to two e-mail reminders were sent to non-responders.

Appraisers who assessed at least one MSF report were e-mailed a personalised record (December 2013) showing how their own assessments compared with those of other appraisers. The appraisal lead encouraged appraisers to reflect on the training exercise and their personalised record as part of their annual quality assurance review, and to discuss learning points in their local appraiser support group.

Statistical analysis

The appraisers who used the online training resource were described in terms of their sex, age, ethnic origin, region of primary medical qualification (PMQ), and appraisal experience. The characteristics of appraisers who assessed at least one MSF report ('participants') were compared with those who registered but did not assess any reports ('non-participants') using χ^2 tests for categorical variables and Mann–Whitney U tests for continuous variables.

For each MSF report, the frequency distribution of responses was described on the six assessment items and, for the three ordinal-scale items (overall assessment, concerns, and acceptability), the mode, mean, and standard deviation (SD) of the ratings were calculated.

For appraisers who assessed all eight MSF reports, an 'agreement score' was calculated by summing the differences

Table 2. Characteristics of participating and non-participating appraisers

		Participating appraisers[a] (*N* = 101), *n* (%)	Non-participating appraisers[b] (*N* = 45), *n* (%)	Statistical tests
Sex	Male	52 (51)	32 (71)	χ^2 (df1) = 4.908; *P* = 0.03
	Female	49 (48)	13 (29)	
Age group, years	30–39	10 (10)	3 (7)	χ^2 (df3) = 3.931; *P* = 0.27
	40–49	35 (35)	10 (22)	
	50–59	43 (43)	26 (58)	
	≥60	9 (9)	6 (13)	
	Missing	4 (4)	0 (0)	
Ethnic group	White	86 (85)	33 (73)	χ^2 (df1[c]) = 6.144; *P* = 0.02
	Mixed	0 (0)	1 (2)	
	Asian or Asian British	9 (9)	10 (22)	
	Chinese or other group	1 (1)	1 (2)	
	Missing	5 (5)	0 (0)	
Region of primary medical qualification	UK	86 (85)	39 (87)	χ^2 (df1c) = 0.116; *P* = 0.78
	European Economic Area	2 (2)	2 (4)	
	South Asia	7 (7)	4 (9)	
	Other	2 (2)	0 (0)	
	Missing	4 (4)	0 (0)	
Length of experience as medical appraiser, years		Median = 7.0 LQ = 3; UQ = 10 (range 0–38)	Median = 8.0 LQ = 3; UQ = 10 (range 0–15)	(Mann–Whitney *U* = 2039.50, *P* = 0.47).

[a]All participating appraisers who assessed at least one MSF report (excluding three appraisers who participated in preliminary user-testing in July–October 2012). [b]Non-participating appraisers registered to use the online training resource but submitted no MSF report assessments. [c]χ^2 test for ethnic group compared white versus other ethnic groups; χ^2 test for region of primary medical qualification (PMQ). PMQ compared UK versus other PMQ regions. df = degrees of freedom. LQ = lower quartile. UQ = upper quartile.

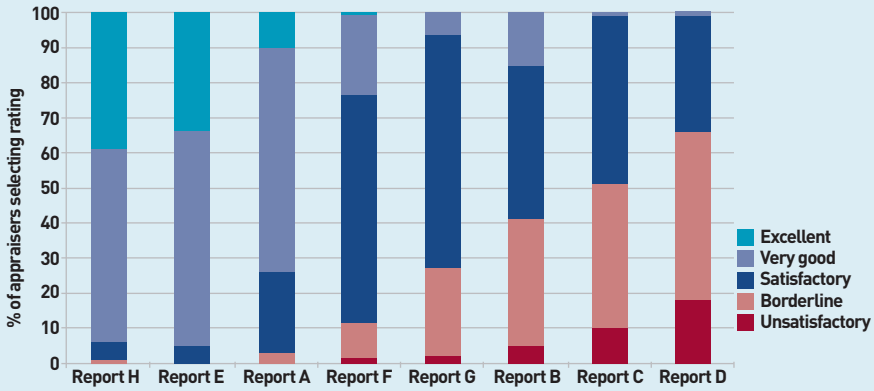

Figure 1. Distribution of appraisers' evaluations of feedback reports A to H (overall assessment ratings). Notes on grouping of reports: (i) Reports A, E, H, and F: doctors' scores on the Patient Questionnaire (PQ) and the Colleague Questionnaire (CQ) were not statistically outlying in previous piloting of the GMC questionnaires. (ii) Reports G, B, and C: doctors' scores on one questionnaire (either the PQ or the CQ) were statistically outlying in previous piloting of the GMC questionnaires. (iii) Report D: doctor's scores on both questionnaires (the PQ and the CQ) would have been statistically outlying in previous piloting of the GMC questionnaires.

between their overall assessment rating and the modal rating of all appraisers on each of the reports. Negative agreement scores were indicative of hawk-like tendencies (on average, rating reports less favourably than peers), whereas positive scores were indicative of dove-like tendencies (on average, rating reports more favourably than peers).

The distribution of these agreement scores was described and an analysis of variance (ANOVA) was conducted to explore the effects of sex, age (four categories), ethnic group (two categories: white, other), PMQ (two categories: UK, other), and years of appraiser experience as predictors of hawk–dove-like tendencies. *P*-values of less than 0.05 were regarded as statistically significant.

RESULTS

Participants

In total, 146/235 (62%) appraisers registered to use the online training resource, of whom 101/146 (69%) assessed at least one MSF report and 86/146 (59%) assessed all eight reports.

Table 2 describes the characteristics of participating and non-participating appraisers. Non-participants were more likely than participants to be male (*P* = 0.03) and from non-white ethnic groups (*P* = 0.02). However, the two groups were similar in terms of age, region of PMQ, and experience as an appraiser (all *P* >0.05).

MSF report assessments

Participants' overall assessments of reports A to H are summarised in Figure 1. Detailed data on the distribution of their responses on all three evaluative scales appear in Appendix 1. The pattern of modal ratings (Table 3) suggests that appraisers were broadly able to detect variation in GPs' MSF results.

Most reports (7/8) received a modal overall assessment of 'Satisfactory' or higher; only report D had a mode assessment of 'Borderline'. Mean concern ratings about GP performance were highest for reports D, C, G, and B. More than half of appraisers reported 'significant concerns' about the GP performance reflected in report D (statistically outlying on PQ and CQ) and in report C (outlying on CQ only). A similar proportion reported 'minor concerns' about the doctors' performance reflected in reports G (outlying on CQ) and B (outlying on PQ). Doctor performance was rated by most appraisers as being 'clearly acceptable' or 'probably acceptable' for all reports. One-quarter to one-third of appraisers, however, rated the performance of the doctors assessed in reports C and D as 'probably unacceptable' or 'clearly unacceptable'.

Given the formative purpose of MSF, appraisers appeared to recommend reasonable actions (Table 4) in the form of repeating one or both surveys, and the inclusion of training in the doctor's PDP. Additional actions were suggested by one-quarter of appraisers in response to report

Table 3. Appraisers' modal evaluations of feedback reports A to H with reports grouped by the pattern of doctor's feedback scores

Report	Pattern of doctor's feedback scores	Evaluative item, *n*/*N* (% participants)		
		Overall assessment of report	Level of concern about doctor's performance	Acceptability of doctor's performance
A	Fell within 'normal distribution' of scores on PQ and on CQ	Very good (64/100; 64)	Not at all (58/100; 58)	Clearly acceptable (73/100; 73)
E		Very good (54/88; 61)	Not at all (76/88; 86)	Clearly acceptable (86/88; 98)
H		Very good (48/87; 55)	Not at all (64/87; 74)	Clearly acceptable (80/87; 92)
F		Satisfactory (58/89; 65)	Minor only (62/89; 70)	Probably acceptable (47/89; 53)
B	Outlier[a] on PQ only	Satisfactory (42/95; 44)	Minor only (44/95; 46)	Probably acceptable (58/95; 61)
C	Outlier[a] on CQ only	Satisfactory (45/93; 48)	Significant (48/93; 52)	Probably acceptable (63/93; 68)
G	Outlier[a] on CQ only	Satisfactory (58/87; 67)	Minor only (55/87; 63)	Probably acceptable (63/87; 72)
D	Outlier[a] on PQ and on CQ	Borderline (44/91; 48)	Significant (62/91; 68)	Probably acceptable (57/91; 63)

[a]Outlying Patient Questionnaire (PQ) or Colleague Questionnaire (CQ) overall scores were >1.96 standard deviations below the mean PQ or CQ overall score (standardised Z score ≤–1.96) calculated for all doctors who participated in GMC questionnaire pilot work.[3]

D, most commonly recommending that the appraiser sought advice from a GP tutor or appraisal supervisor, discussed mental wellbeing and stress management with the doctor, and explored the doctor's insight into their communication skills. Few appraisers recommended referring GPs to their responsible officer for further review

Table 4. Appraisers' suggested actions for feedback reports A to H: distribution of responses

	Report, *n* (%)							
	A	B	C	D	E	F	G	H
	N = 100	*N* = 95	*N* = 93	*N* = 91	*N* = 88	*N* = 89	*N* = 87	*N* = 87
Repeating patient and colleague surveys								
No need to repeat either survey	**90 (90)**	**45 (47)**	**43 (46)**	17 (19)	**87 (99)**	**57 (64)**	**46 (53)**	**85 (98)**
Repeat the patient survey only	0 (0)	**45 (47)**	3 (3)	33 (36)	1 (1)	12 (13)	6 (7)	2 (2)
Repeat the colleague survey only	7 (7)	0 (0)	17 (18)	2 (2)	0 (0)	2 (2)	20 (23)	0 (0)
Repeat both surveys	3 (3)	5 (5)	30 (32)	**39 (43)**	0 (0)	18 (20)	15 (17)	0 (0)
Personal development plan (PDP) actions								
No specific PDP action(s) or training	43 (43)	12 (13)	1 (1)	1 (1)	**65 (74)**	18 (20)	5 (6)	**63 (72)**
Encourage to include training in PDP	**55 (55)**	**57 (60)**	**49 (53)**	40 (44)	23 (26)	**61 (69)**	**62 (71)**	24 (28)
Mandate to include training in PDP	2 (2)	26 (27)	43 (46)	**50 (55)**	0 (0)	10 (11)	20 (23)	0 (0)
Other recommended action(s)[a]								
No other action needed	70 (70)	32 (34)	7 (8)	2 (2)	**77 (88)**	**42 (47)**	24 (28)	**63 (72)**
Review PDP actions at next appraisal	30 (30)	**60 (63)**	**80 (86)**	**73 (80)**	8 (9)	**42 (47)**	**60 (69)**	17 (20)
Refer to the responsible officer	0 (0)	1 (1)	4 (4)	6 (7)	0 (0)	0 (0)	1 (1)	0 (0)
Other action recommended	0 (0)	8 (8)	13 (14)	25 (27)	3 (3)	9 (10)	10 (11)	7 (8)

[a]Percentages may add up to more than 100% because appraisers could select more than one action. PDP = professional development plan. Emboldened figures represent modal response(s).

(1%, 5%, and 7% for reports B, C, and D, respectively).

Hawk–dove effects

Agreement scores, reflecting the difference between individual appraisers' assessment ratings and the modal rating for the eight MSF reports, ranged from –7 to +7 (mean agreement score 0.49, SD 3.01). An agreement score of –7 would indicate a hawk-like appraiser who may, for example, have rated seven of the eight reports at 1 point below the modal rating (for example, 'Borderline' rather than 'Satisfactory') and agreed with the modal score on just one report. Conversely, an agreement score of +7 would indicate a dove-like appraiser who may also have given the modal rating on one report but rated the other seven reports at 1 point higher than the mode. Hawk-like tendencies were more common (44/86, 51% appraisers with a negative agreement score) than dove-like tendencies (29/86, 34% appraisers with a positive agreement score). Despite this, the mean agreement score was positive, indicating that the dove-like raters tended to deviate more from the modal rating than did the hawk-like raters.

Age was a significant predictor of hawk/dove-like tendencies, with older appraisers rating the MSF reports more favourably than younger appraisers (B = 0.129, $P = 0.01$). Sex, ethnic origin, PMQ, and years as an appraiser were not, however, significant predictors of hawk/dove-like tendencies.

DISCUSSION

Summary

Despite the complexity of information in the featured MSF reports, appraisers' assessments suggested that they could detect variations in MSF score patterns. For each report, there was broad consensus about the level of concern and acceptability of the GP's performance (based on the information in their MSF report) and about actions that could be discussed in the appraisal meeting. Appraisers varied, however, in their tendency to be more stringent or lenient in their assessment of MSF reports relative to their peers. In particular, there was some evidence that older appraisers may be more lenient than younger appraisers in this regard.

Strengths and limitations

Seven of the eight MSF reports had been issued to practising GPs,[3] and therefore appraisers assessed realistic MSF information. The design of the online resource meant that appraisers could review reports over several sessions to fit around their work

schedule. Appraisers' judgements about a doctor's performance were made solely on the basis of an MSF report, however, without access to the doctor's other supporting information, or knowledge of the doctor's reaction to their feedback, which would occur in a real appraisal context.

A number of MSF tools that include different items, scales, and reporting formats are available to doctors. The present study focused only on the interpretation of MSF reports derived from the GMC Patient and Colleague Questionnaires.

GP appraisers were drawn from one region of the UK, which limits generalisability to other regions and contexts in which MSF is used. Relatively small numbers of appraisers participated ($N = 101$), and participants were more likely to be female and from white ethnic backgrounds. The present findings should, therefore, be regarded as preliminary and interpreted cautiously until replicated with other appraiser samples.

With regard to hawk–dove effects in interpreting MSF, limited demographic information was collected about participating appraisers, and other factors not addressed in this study may be associated with the observed variation in stringency or leniency.

Comparison with existing literature

The present observation that individual appraisers may vary in the leniency of their assessments of MSF reports is in line with hawk–dove effects observed in relation to other practice-based assessments.[12–17] Previous research has identified demographic characteristics of assessors that may be associated with variations in leniency (such as ethnic origin and experience).[14,16] The present study has identified appraiser age, but not length of experience as an appraiser, as a potential predictor of greater leniency in interpreting MSF reports. Appraisers from non-white ethnic backgrounds were under-represented in the present sample and this may account for the absence of an observed effect of ethnic origin on leniency in this study.

Implications for research and practice

The present study suggests that appraisers can detect variation in the pattern of GPs' MSF scores and recommend appropriate actions based on a review of complex MSF information. Furthermore, as a group, the appraisers were reasonably consistent in their interpretations of each doctor's MSF results. This observation should be reassuring for GPs and appraisers, as well as for appraisal leads, responsible officers, and designated bodies[11] who have responsibility for quality assurance of the appraisal processes.

Individual differences in leniency were observed in appraisers' interpretations of MSF, which may be linked to the appraiser's age. GPs' experiences of reflecting and acting on MSF within their appraisal may therefore vary according to the age of their appraiser. The extent to which this proves problematic in real-life practice has yet to be established. Similarly, the need for organisations to take steps to attenuate appraiser differences in leniency around MSF requires further consideration. This could include the use of training packages utilising standardised reports, such as those described in this study. Future development work could evaluate appraisers' views of the online training resource and determine how it could be improved by seeking feedback from appraisers who assess all eight MSF reports as well as those who assess fewer reports.

Research employing qualitative or cognitive interviewing methods may explore how appraisers arrive at judgements about a doctor's performance based on MSF reports, and which aspects of the available MSF information influence their interpretations. Further study of hawk–dove effects in this context could identify why such differences exist, how appraisers view their own level of stringency, and whether these effects change after using the training resource or change with increasing experience of interpreting MSF in the context of 'real-world' appraisal.

Funding

The work was supported by a research grant from the General Medical Council (GMC) and funds allocated by Health Education North East for appraiser training.

Ethical approval

The project represented the development and piloting of a local training resource for NHS appraisers and therefore did not require NHS ethics approval.

Provenance

Freely submitted; externally peer reviewed.

Competing interests

John Campbell was an advisor to the GMC during the development of the GMC patient and colleague questionnaires (2005–2011) and received only direct costs associated with presentation of that work. The other authors have declared no competing interests.

Acknowledgements

The authors thank the appraisers who piloted and helped to refine the online training resource materials in the various development phases. They also thank Zac Gribble and Sally Holden (formerly of the E-Learning Support Team, Peninsula College of Medicine and Dentistry), who provided technical expertise during the development of the initial prototype of the online training resource. A summary of this work was presented as an 'elevator pitch' at the Society for Academic Primary Care (SAPC) National Meeting in Edinburgh, Scotland (11 July 2014).

Discuss this article

Contribute and read comments about this article: **bjgp.org/letters**

REFERENCES

1. General Medical Council. *Good medical practice*. Manchester: GMC, 2013.
2. General Medical Council. *Supporting information for appraisal and revalidation*. Manchester: GMC, 2012.
3. 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.
4. Campbell JL, Roberts M, Wright C, *et al*. Factors associated with variability in the assessment of UK doctors' professionalism: analysis of survey results. *BMJ* 2011; **343:** d6212.
5. Hill JJ, Asprey A, Richards SH, Campbell JL. Multisource feedback questionnaires in appraisal and for revalidation: a qualitative study in UK general practice. *Br J Gen Pract* 2012; DOI: 10.3399/bjgp12X641429.
6. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ* 2010; **341:** c5064.
7. 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.
8. Sargeant J, Mann K, Sinclair D, *et al*. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv Health Sci Educ Theory Pract* 2008; **13(3):** 275–288.
9. NHS Revalidation Support Team. *Quality assurance of medical appraisers: recruitment, training, support and review of medical appraisers in England*. London: NHS Revalidation Support Team, 2013.
10. NHS Revalidation Support Team. *Medical appraisal guide: a guide to medical appraisal for revalidation in England*. London: NHS Revalidation Support Team, 2013.
11. Royal College of General Practitioners. *The principles of GP appraisal for revalidation*. London: RCGP, 2014.
12. Bartman I, Roy M, Smee S. *Catching the hawks and doves: a method for identifying extreme examiners on objective structured clinical examinations (technical report)*. Ottawa, ON: Medical Council of Canada, 2011.
13. Harasym PH, Woloschuck W, Cunning L. Undesired variance due to examiner stringency/leniency effecct in communication skill scores in OSCEs. *Adv Health Sci Educ Theory Pract* 2008; **13(5):** 617–632.
14. McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk–dove effect') in the MRCP (UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Med Educ* 2006; **6:** 42.
15. Roberts C, Rothnie I, Zoanetti N. Should candidate scores be adjusted for interviewer stringency or leniency in multiple mini-interviews? *Med Educ* 2010; **44(7):** 690–698.
16. McManus IC, Elder AT, Dacre J. Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP(UK) PACES and nPACES examinations. *BMC Med Educ* 2013; **13:** 103.
17. Bartman I, Smee S, Roy M. A method for identifying extreme OSCE examiners. *Clin Teach* 2013; **10(1):** 27–31.

Appendix 1. Appraisers' evaluations of feedback reports A to H: distribution and mean ratings (for overall assessment of report; concerns about doctor's performance; and acceptability of doctor's performance)

	Report, *n* (%)							
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
	***N*= 100**	***N*= 95**	***N*= 93**	***N*= 91**	***N*= 88**	***N*= 89**	***N*= 87**	***N*= 87**
Overall assessment of report								
Excellent (5)	10 (10)	0 (0)	0 (0)	0 (0)	30 (34)	1 (1)	0 (0)	34 (39)
Very good (4)	**64 (64)**	14 (15)	1 (1)	1 (1)	**54 (61)**	20 (22)	5 (6)	**48 (55)**
Satisfactory (3)	23 (23)	**42 (44)**	**45 (48)**	30 (33)	4 (5)	**58 (65)**	**58 (67)**	4 (5)
Borderline (2)	3 (3)	34 (36)	38 (41)	**44 (48)**	0 (0)	9 (10)	22 (25)	1 (1)
Unsatisfactory (1)	0 (0)	5 (5)	9 (10)	16 (18)	0 (0)	1 (1)	2 (2)	0 (0)
Mean rating (SD)	*3.8 (0.65)*	*2.7 (0.79)*	*2.4 (0.68)*	*2.2 (0.72)*	*4.3 (0.55)*	*3.1 (0.64)*	*2.8 (0.59)*	*4.3 (0.62)*
Concerns about doctor's performance								
Not at all concerned (1)	**58 (58)**	9 (9)	0 (0)	0 (0)	**76 (86)**	17 (19)	5 (6)	**64 (74)**
Minor concerns only (2)	40 (40)	**44 (46)**	43 (46)	27 (30)	12 (14)	**62 (70)**	**55 (63)**	23 (26)
Significant concerns (3)	2 (2)	42 (44)	**48 (52)**	**62 (68)**	0 (0)	10 (11)	27 (31)	0 (0)
Extremely concerned (4)	0 (0)	0 (0)	2 (2)	2 (2)	0 (0)	0 (0)	0 (0)	0 (0)
Mean rating (SD)	*1.4 (0.54)*	*2.3 (0.65)*	*2.6 (0.54)*	*2.7 (0.50)*	*1.1 (0.34)*	*1.9 (0.55)*	*2.2 (0.55)*	*1.3 (0.44)*
Acceptability of doctor's performance								
Clearly acceptable (4)	**73 (73)**	22 (23)	7 (8)	3 (3)	**86 (98)**	39 (44)	15 (17)	**80 (92)**
Probably acceptable (3)	25 (25)	**58 (61)**	**63 (68)**	**57 (63)**	2 (2)	**47 (53)**	**63 (72)**	7 (8)
Probably unacceptable (2)	2 (2)	15 (16)	20 (22)	27 (30)	0 (0)	3 (3)	9 (10)	0 (0)
Clearly unacceptable (1)	0 (0)	0 (0)	3 (3)	4 (4)	0 (0)	0 (0)	0 (0)	0 (0)
Mean (SD) rating	*3.7 (0.50)*	*3.1 (0.62)*	*2.8 (0.61)*	*2.6 (0.62)*	*4.0 (0.15)*	*3.4 (0.56)*	*3.1 (0.52)*	*3.9 (0.27)*

Emboldened figures represent modal response(s). SD = standard deviation. Overall scores, which are emboldened in the table were defined as those lying >1.96 standard deviations below the mean Patient Questionnaire or Colleague Questionnaire overall score (standardised Z score < −1.96) calculated for all doctors who participated in GMC questionnaire pilot work.[3] Means and SD for each number on each evaluation scale were calculated by scoring the ordinal-response scale items as indicated in column one, and are italicised on the table.