Using a 'peer assessment questionnaire' in primary medical care

Glyn Elwyn, Malcolm Lewis, Richard Evans and Hayley Hutchings

ABSTRACT

Background

Periodic assessment of clinician performance or 'revalidation' is being actively considered to reassure the public that doctors are 'up to date and fit to practice'. There is, therefore, increasing interest in how to assess individual clinician performance in a valid and reliable way. The use of peer assessment questionnaires is one of the methods being considered and investigated by the General Medical Council in the UK.

Aim

To test the feasibility of using a peer assessment questionnaire in a primary care setting, and consider the related issues of validity and reliability and compare the results to previous studies.

Design

Cross-sectional survey in a volunteer sample.

Setting

General practice in the UK.

Method

GPs who volunteered to take part in an evaluation of a pilot appraisal implementation scheme were recruited by appraisers. These volunteers (GP subjects) chose 15 colleagues to complete a 'peer assessment' questionnaire that asked peers to make judgements about their clinical skills and other characteristics, such as 'compassion', 'integrity' and 'responsibility'.

Results

Of the 207 practitioners that agreed to be appraised, 113 completed the optional task of implementing the peer questionnaire. Of the 1271 raters, 1189 provided data about their roles and 33.6% of these were GPs. The data revealed significant levels of items where peers were 'unable to evaluate' the issues posed in the questionnaire (ranging from 13.7–1.8%). These rates differed from those obtained in studies based in the US where mean scores were slightly higher. Although the overall results are broadly similar to those previously obtained, there are sufficient differences to suggest that there are contextual issues influencing the interpretation of the items and therefore the scoring process.

Conclusion

The volunteer sample in this study found no major obstacles to the implementation of the peer assessment questionnaire. While it is not possible to generalise from this selected volunteer sample, the use of peer assessment questionnaires appears feasible and may be acceptable to clinical practitioners. However, concern remains about the validity of such instruments and that their development did not fully consider issues of procedural justice or whether the overall purpose of the tools was to be formative, summative, or both.

Keywords

employee performance appraisal; peer assessment questionnaire; revalidation.

INTRODUCTION

Clinicians, in their less guarded moments, are known to confide with each other that the 'bad apples', the poor performers, among them are 'known' to all, but that it is difficult to obtain confirmatory evidence of such deficit. Perhaps a system of assessment based on 'scores' given by one's professional colleagues could provide an indication of the variation that exists in practice. In addition to the more general aim of assessing performance for formative reasons, interest is increasing in the concept of periodic assessment of performance, which, in the UK, goes by the name of 'revalidation'. The stated aim of revalidation in the UK is to 'ensure that patients have the confidence that licensed doctors are up to date and fit to practice'.[1] This definition by its nature implies a summative judgment is being made and that, as a consequence, some doctors will be found wanting. Controversy continues in the summer of 2005 about how revalidation will be conducted.[2] Initial proposals for a standards-based 'folder of evidence' were put aside by the General Medical Council (GMC), so that by 2001 they were advising that revalidation should be based on five annual appraisals, augmented by clinical governance reviews.[2] However, in December 2004, Dame Janet Smith's fifth Shipman Inquiry report led to a decision, that revalidation, as

***G Elwyn**, BA, MSc, PhD, FRCGP, research professor, Centre for Health Sciences Research, School of Medicine; **M Lewis**, LLM, FRCGP, ITLM, director of postgraduate education for general practice, Department of Postgraduate Education for General Practice, Wales College of Medicine, Cardiff University. **R Evans**, MA, MRCGP, ITLM senior clinical tutor; **H Hutchings**, BSc, PhD, lecturer, Primary Care Group, School of Medicine, Swansea University.*

Address for correspondence

Glyn Elwyn, Research Professor, Centre for Health Sciences Research, School of Medicine, Cardiff University, 56 Park Place, Cardiff CF10 3AT. E-mail: elwyng@cardiff.ac.uk

Submitted: 2 March 2005; **Editor's response:** 26 May 2005; **final acceptance:** 15 July 2005.

©*British Journal of General Practice* 2005; **55:** 690–695.

formulated, should be postponed, pending a review by the Chief Medical Officer.[3]

Given the interest in how to assess the performance of doctors, there has been a search for structured systems for gathering information about the performance of doctors, specifically that clinicians are able to demonstrate that their practice is based on the principles of *Good Medical Practice*.[4] One of the requirements of *Good Medical Practice* is the provision of evidence of good working relationships with colleagues, with patients and information on aspects of health and probity, yet there is lack of clarity about how to collect this information.

If doctors are able to make judgements about which doctors perform well (and by corollary, which do not), then the possibility has been entertained that instruments might be designed to collect this information in a systematic and rigorous way. This data could then be used for assessment purposes, be that for formative use, such as in appraisal, or for a summative judgement about fitness to practice. A review of these tools has been recently published which raised a series of issues for discussion.[5] It is known that the GMC has developed a questionnaire for this purpose and has commissioned an evaluation of the instrument's validity, reliability and acceptability and is presumably considering its use in the revalidation process. There is, therefore, a need to consider the acceptability and the implications of these measurement methods in more detail.

Evaluative judgement of clinician colleagues using a formalised peer rating measure began to be used in the US during the 1980s and in Canada in the 1990s.[6–9] There was a perceived need for increasing accountability to regulatory bodies and patients, accompanied by a shift in thinking about what constituted a full spectrum of clinical competence to include non-cognitive 'humanistic' qualities, such as compassion, integrity, responsibility, respect and interpersonal communication skills. The need to ensure that these areas were 'demonstrated' and the inability of conventional examinations to evaluate these qualities aroused interest in developing and refining a reliable means of peer evaluation. Ramsey first demonstrated the potential of a Physician Associate Rating (PAR)[10] in the US as an evaluative tool in addressing the wider definition of clinical competence set by the American Board of Internal Medicine (ABIM),[11] and determined its psychometric properties for US physicians practising as general internists, outside the research setting.[11–12] The PAR is now part of the 'patient and peer assessment module' of the ABIM's Continuous Professional Development programme. In Canada the Physician Achievement Review has the aim of quality improvement using a supportive educational process. Their rating instrument here, the Peer Assessment Questionnaire (PAQ) was derived from a grid of performance attributes, developed by pre-piloting and focus groups, and evaluated for its psychometric properties. Both the US and Canadian instruments are used for quality improvement initiatives, supporting and educating clinicians. The Canadian instrument seems to have a more robust developmental pathway that involved qualitative studies (focus groups), although unpublished. The US instrument is based on recommendations from two reports.[11,13] The instrument's performance in a primary care setting has not been demonstrated in the US because studies were conducted on general internists and specialists in secondary care, whereas in Canada the majority (80%) of those using the PAQ were general family physicians. Given the paucity of information, it is important to study how feasible and applicable these instruments can be to primary care in other settings, such as the UK.

An appraisal scheme was being piloted and evaluated.[14] In the context of this evaluation, it was decided to ask those who volunteered to take part in the pilot appraisal scheme to also use a peer questionnaire, an adaptation based on the tool that Ramsey had developed in the US on behalf of the ABIM.[12] The aim was to test the feasibility of using a peer assessment questionnaire in a primary care setting, and, by so doing, consider validity and reliability and compare the results to previous findings.

How this fits in

There are a handful of studies which suggest that it is possible for peers to accurately assess each other using questionnaires that specify different aspects of clinical practice. Although there seems to be evidence that peers are willing to undertake this sort of assessment and that there is a consistency of scores across studies, concerns remain about the validity of these tools and a lack of consensus about how the scores should be used. This study is the first to use these tools in a UK setting.

METHOD

The questionnaire

The original ABIM peer questionnaire is available for inspection on the ABIM website (www.abim.org/). The instrument was reviewed and modified by one of the authors for use in a UK primary care setting, following the principle of retaining as much as was feasible of the original tool. It was felt that changes should be kept to a minimum and only introduced in order to reflect the context of use in UK primary care. The original document used by the ABIM contained 11 questions. The following changes were made: items that referred to hospital outpatient and inpatient

settings were changed to general practice surgeries and 'ambulatory care skills in the outpatient setting' were changed to 'primary care skills in the surgery setting'. Question 9 of the questionnaire relating to care of hospitalised patients was excluded, being inappropriate for the majority of GPs, who do not have direct responsibility for hospitalised patients. It would be reasonable to conclude in numerical terms that $^{10}/_{11}$ of the original questionnaire was used and that 90% of the questionnaire used was unchanged, albeit with modifications made to reflect a primary care UK context.

Sample of GPs

In May 2001, details about 20 appraiser posts were circulated to 1800 principals and 160 non-principals in Wales. The aim was to recruit appraisers who represented a cross section of NHS GPs in terms of age, sex, geographical location, size of practice, ethnic backgrounds and contractual status (principals and non-principals). Selection procedures excluded clinicians who had prior educational experience. Further details are published elsewhere.[14] To recruit clinicians who were willing to undergo the appraisal process, all GPs in Wales were informed about the formative aim of the exercise and the proposed potential integration with revalidation. Contact details were provided so that individuals could make direct approaches to the appraisers of their choice.

Implementation of the peer-rating questionnaire

The peer questionnaires were sent by post to the appraisers. Each appraiser asked the GPs who had volunteered to take part in the pilot appraisal scheme to use the peer assessment instrument. It was emphasised that the process was voluntary and that the aim was to assess the feasibility of using these instruments as well as compare the results with those of published studies. As in previous studies, the GPs were advised to ask 15 colleagues to complete the questionnaires. Colleagues were defined as GPs in the same partnership or working in the same local setting, or any other medical, surgical, nursing or administrative colleagues who they felt could make the judgements required on the questionnaire. No data was collected on the numbers of colleagues approached. Those completing questionnaires (the peers assessors) were told that their comments would remain anonymous: questionnaires were received back at the practice before being returned to the appraisers. Appraisers then returned the peer questionnaires for analysis.

Analysis

Each GP-subject was evaluated by a number of peers using the distributed questionnaires. Each rater was asked to complete a questionnaire composed of 10 items that related to the GPs ability (competence) and provide details regarding their professional relationship to the GP-subject, such as employee or professional partner or receiver of patient referrals (Supplementary Figure 1). For each of the 10 items a 9-point Likert scale was used, ranging from 1 (the practitioner was the worst GP the rater had ever worked with) to 9 (the practitioner was the best they had ever worked with). If the rater had insufficient contact to evaluate the GP on a particular category, they were asked to indicate that they were unable to complete that item. Ratings for each GP-subject were used to calculate a mean rating for each of the 10 categories, score ranges were also calculated. Mean UK GP ratings in each of the 10 items were compared with the results for board certified US internists studied by Ramsay *et al*.[7] Scores were calculated using all available ratings. We felt it important to analyse the potential difference in score ranges that are achieved when a lower number of ratings per GP-subject were achieved, so, therefore, sensitivity analyses conducted for ratings by a number greater than 5, 7 and 10 peers.

RESULTS

The characteristics of raters are provided in Table 1. Those raters who were themselves GPs were based in a range of practices, 49.5% in large group practices, 33.4% in small group practices, 3.5% in single-handed practices (13.6% did not complete information about practice size); 92 raters did not complete this category. In terms of professional relationship to the GP-subject, the highest percentage of raters (39.9%) were employees of the index practitioner, a clear source of possible bias. A further 21.3% were in partnership with the GP, 18.8% of the raters had patients referred from the GP and the remainder (20%) had some other professional relationship with the GP; 95 raters did not complete this category. The length of relationship of the raters to the GP was greater than 1 year in the majority of cases (87.6%) indicating the duration of professional

Table 1. Characteristics of the raters.

Number of raters	1271
Role ID provided (%)	1189 (94%)
Male/Female (%)	820/369 (69/31)
Number of GPs (%)	399 (33.6)
Nurses (%)	341 (28.7)
Management (%)	285 (23.9)
Others (%)	
Physicians	3.7
Surgeons	2.9
Others	7.2

Table 2. Proportion of missing or unevaluable cases for each category.

Category	Missing values (%)	UK unevaluable values (%)	Total (%)	US unevaluable cases (%)
Respect	6 (0.5)	32 (2.5)	3	11
Medical knowledge	10 (0.8)	79 (6.2)	7	1.8
Primary care skills (Ambulatory care skills in US version)	10 (0.8)	96 (7.6)	8.4	13.7
Integrity	5 (0.4)	25 (2)	2.4	2.3
Psychosocial aspects	10 (0.8)	130 (10.2)	11	9.1
Management of complex problems	13 (1)	135 (10.6)	11.6	5.4
Compassion	2 (0.2)	44 (3.5)	3.7	7.3
Responsibility	2 (0.2)	30 (2.4)	2.6	2.2
Problem solving	8 (0.6)	66 (5.2)	5.8	3.2
Clinical skills	11 (0.9)	99 (7.8)	8.7	1.8

contact on which the assessment was based. A further 7.7% claimed to have known the GP for between 6 months and 1 year, and 4% of the raters had known the GP for less than 6 months. A further 0.6% of raters ticked the 'not-applicable' category; 82 of the raters did not complete this category. This study was not designed to study the impact of organisational size (cluster) on peer ratings, employer–employee power dynamic or professional relationship duration to GP-subject, but this may need to be considered in future evaluations.

Completion rates and missing data

A total of 1271 questionnaires were completed relating to 113 GP-subjects. The number of ratings for each GP ranged from 2–23 with a median number of 12. Table 2 lists the missing and unevaluable items from the questionnaire dataset.

Among the GP peer responses, the percentage of 'unable to evaluate' responses was analysed for each individual rating category. There was considerable variation in the percentage of peers unable to evaluate a GP in different categories (Table 2). The percentages of 'unable to evaluate' were highest for the management of complex problems (10.6%), psychosocial aspects (10.2%), clinical skills (7.8%) and primary care skills (7.6%). The percentages of peers 'unable to evaluate' were lowest in the categories of integrity (2%), responsibility (2.4%), respect (2.5%) and compassion (3.5%). In the remaining two categories, the percentages were moderately low (5.2% for problem solving and 6.2% for medical knowledge). These results indicate that raters found that the areas where they were unable to evaluate were those that involved an assessment of their colleagues' clinical performance rather than the broader aspects such as overall 'integrity'. The 'unevaluable' results are different from the results of the Ramsay study (Table 2). In the US study, the highest 'unable to evaluate' categories were primary care skills (13.7%), respect (11%), psychosocial aspects (9.1%) and compassion (7.3%).

The peer questionnaires ratings for the doctors were largely similar to those obtained in the US studies (Table 3), although important differences are noted. The mean scores for all the included items were slightly higher than those obtained in the US and higher maximum scores were observed. It's

Table 3. Mean ratings for 113 practitioners who were rated by 2 or more peers.

Rating category	Range of mean ratings	Mean rating	SD mean rating
Respect	4.88–9.00 6.08–8.91[a]	8.09 7.78[a]	0.59 0.5[a]
Medical knowledge	6.75–9.00 6.20–8.62[a]	8.20 7.63[a]	0.44 0.57[a]
Primary care skills (UK) Ambulatory care skills (US)	6.88–8.90 6.11–8.71[a]	8.22 7.67[a]	0.39 0.5[a]
Integrity	7.08–9.00 6.18–9.00[a]	8.37 8.11[a]	0.38 0.43[a]
Psychosocial aspects	6.45–8.86 5.75–8.73[a]	0.38 7.57[a]	0.49 0.55[a]
Management of complex problems	5.50–9.00 5.87–8.67[a]	8.10 7.58[a]	0.48 0.62[a]
Compassion	6.43–9.00 5.77–8.82[a]	8.18 7.7[a]	0.51 0.56[a]
Responsibility	6.29–9.00 6.18–9.00[a]	8.35 7.98	0.45 0.46[a]
Problem solving	5.86–9.00 5.93–8.75[a]	8.11 7.7[a]	0.48 0.54[a]
Clinical skills	6.13–9.00 6.21–8.67[a]	8.18 7.71[a]	0.46 0.53[a]
Management of (US) hospitalised patients	6.13–8.71[a]	7.71[a]	0.53[a]

[a]*Results from Ramsay's US study.*[7] *SD = standard deviation.*

possible that the lack of an anonymous system for rating, as compared to the coded telephone scoring method in US, might have led to more generous scores being given, coupled with the potential influence of a wider range of professions involved in the rating and the intra-organisational (and often employee status) of the rater. Nevertheless, the data were distributed on a normal curve, with a slight predominance of values towards the upper end. A sensitivity analysis was performed to assess if the mean scores changed if there were greater than 5, 7 or 10 ratings. No significant differences were found. Some of the items do show a more compressed range when greater than 10 raters provide scorings.

DISCUSSION

Summary of main findings

It was clear during this pilot scheme of a voluntary appraisal system in general practice that about half (54%) of the GPs were willing to undertake the process of asking peers to use a questionnaire to make judgements on their clinical practice. Similarly the peers recruited by the GP-subjects were willing to complete the ratings, although there are important areas of clinical practice where significant percentages are unable to evaluate questionnaire items. The results are broadly comparable to the scores obtained when the instrument was used in the US. In summary, we draw the conclusion that the use of peer questionnaires in this selected volunteer sample is feasible to both GP-subjects and their peers. Whether this feasibility can be taken to also indicate that this form of assessment is acceptable to a broader cohort of practitioners needs to be investigated.

However, we advise caution. The study was based within an appraisal context and the implicit understanding was that these questionnaires were being tested in a setting that had formative aims. The response of GP subjects and raters might have been different if they had been asked to complete these questionnaires knowing that their scores would contribute to a summative judgment (a pass or fail test). There are also concerns that arise from the results themselves. The data show high levels of 'inability to evaluate'; levels that range from 1.8 to 13.7%, with a high degree of variation between the UK and US settings, suggesting that there are differences influencing the scoring process related to either the raters or their interpretation of the items. If the eventual aim is to use peer ratings to contribute to summative judgements about fitness to practice, then the lack of confidence about the inability to rate the areas dealing with clinical competence, that is, the areas where revalidation has to stand or fall, is problematic. It could be argued that patients can make their own judgments about the humanistic elements (for example, integrity, respect and compassion) but that an episodic assessment has to take care of the technical aspects of practice. The UK results suggest more willingness to rate the 'humanistic' items. Perhaps this reflects that GPs in the UK have a clearer conception of these areas compared to hospital practitioners in the US. It is possible that hospital practitioners witness each other's practice more often, and consult in a more public manner. Most clinical work in UK primary care occurs in the privacy of the consultation room making it difficult to make judgments about practice that is not directly observed.

Strengths and limitations of the study

This is the first study in the UK involving the use of a peer questionnaire and provides evidence that the peer assessment questionnaire system is indeed feasible in a UK primary care setting. It was based, with slight modification, on an existing tool that had established measurement characteristics. This provided the ability for the data generated to be compared to other datasets although caution about the data collection contexts is needed. The 'volunteer-within-selected sample' nature of the subjects needs to be recognised. This group of practitioners are likely to be those most confident to experiment with novel assessment methods. Note that the participants in the Ramsey studies[7] were also volunteers, so we cannot predict whether these instruments would be acceptable to all clinicians. We did not set out to investigate construct or criterion validly and no studies of this nature have been conducted. Another weakness we perceive is the lack of advance rater instruction or preparation and how the scores are to be benchmarked and shared with subjects. The behaviour of raters might be influenced significantly if they know to what extend their anonymity is protected and if detailed advance information were available about the aim and nature of the feedback given to the index-clinician.

Comparison with existing literature

Looking at the wider literature it seems clear that there remains much research to be done on understanding the differences among rater groups.[15,16] In this study, on average, only three of the rater sample are true peers, that is, fellow GPs. It's likely that different rater samples for GP-subjects, composed of secondary care colleagues, may arrive at different judgements reflecting different frames of reference.[17–19] Most of the raters who provided scores in this study worked in the same organisations, contributing to a potential organisational 'cluster' effect, plus a possible bias because scores are potentially attributable. Nevertheless, only

colleagues who work in close proximity could be reasonably expected to have sufficient information to undertake peer ratings. It therefore follows that strict anonymity (using technologies such as remote score transfer) and the definition of a highly specified sample size and composition would be a necessary requirement if scores are to be used for summative judgments. This issue requires further investigation.

Although this study has considered the feasibility of using the peer questionnaire, it is also important to consider the concerns that are emerging about the use of peer assessment questionnaires. While it might be possible to generate 'scores' there remains a worry about the validity of the instruments, the quality of their design and development and lastly, the clarity of their overall purpose.[5] Measuring areas of clinical practice carries with it a responsibility for ensuring a clarity of purpose. This is especially pertinent given the debate about the proposed aims of the revalidation process in the UK.[2,20,21]

Implications for future policy and clinical practice

In the wake of Dame Janet Smith's Shipman Inquiry report, which declared that existing plans for revalidation were not 'fit for purpose',[3] the GMC, having decided to stick to its guns, awaits the Chief Medical Officer's review of revalidation. Meanwhile, the GMC has commissioned an evaluation of a peer assessment tool that it has developed internally — data are awaited. In this context, therefore, this study provides useful evidence to guide policy. We conclude that a peer assessment questionnaire of this nature can indeed be used: instruments of this type have the ability to provide scores at the level of individual practitioners. To date however, the use of these tools has not been performed in contexts where the aims have been made fully explicit. It is one thing to use scores to provide formative feedback in an appraisal-type scheme, although even a formative process of this nature needs considerable thought about how to benchmark scores and arrange a system for supportive feedback. Given the wider concerns about validity, we feel that it is premature to advocate the use of peer assessment questionnaires for summative purposes. It is necessary first to attend to issues of validity, to consider the measurement process issues of sample and anonymity. It is also important to ensure that those who are asked to make judgements in these questionnaires understand the purpose of generating a score, and that they also have sufficient grounds on which to make judgments about both the technical and humanistic components of medical practice. In other words, that the measurement process conforms to the concept known as procedural justice,[22] where the conduct of the assessment is regarded fair and accurate.

Supplementary information

Additional information accompanies this article at http://www.rcgp.org.uk/journal/index.asp

Funding body

Department of Postgraduate Education for General Practice, Cardiff University

Ethics committee

Ethical approval was not sought

Competing interests

None

Acknowledgements

All the appraisers and GPs who volunteered to complete the peer assessment questionnaires are acknowledged.

REFERENCES

1. General Medical Council. *Developing medical regulation: a vision for the future. The GMC's response to the call for ideas by the review of clinical performance and medical regulation.* London: General Medical Council, 2005.
2. Pringle M. *Revalidation of doctors: the credibility challenge. John Fry Fellowship Lecture.* London: The Nuffield Trust, 2005.
3. The Shipman Inquiry. Safeguarding patients: lessons from the past, proposals for the future. London, TSO, 2004. http://www.the-shipman-inquiry.org.uk/fifthreport.asp (accessed 13 Jun 2004).
4. General Medical Council. *Good medical practice.* London: General Medical Council, 2001.
5. Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. *BMJ* 2004; **328(7450):** 1240–1243.
6. Violato C, Marini A, Toews J, *et al.* Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Acad Med* 1997; **72(10 Suppl 1):** S82–S84.
7. Ramsey PG, Carline J, Blank L, Wenrich M. Feasibility of hospital-based use of peer ratings to evaluate the performance of practicing physicians in academic medicine. *Acad Med* 1996; **71:** 364–370.
8. Lipner R, Blank L. The value of patient and peer ratings in recertification. *Acad Med* 2002; **77(10 Suppl):** S64–S66.
9. Hall W, Violato C, Lewkonia R, *et al.* Assessment of physician performance in Alberta: the physician achievement review. *CMAJ* 1999; **161:** 52–57.
10. Ramsey P, Carline J, Inui T, *et al.* Predictive validity of certification by the American Board of Internal Medicine. *Ann Intern Med* 1989; **110:** 719–726.
11. American Board of Internal Medicine. Attributes of the general internist and recommendations for training. *Ann Intern Med* 1977; **86(4):** 472–473.
12. Ramsey P, Wenrich M, Carline J, *et al.* Use of peer ratings to evaluate physician performance. *JAMA* 1993; **269:** 1655–1660.
13. American Board of Internal Medicine. *A guide to awareness and evaluation of humanistic qualities in the internist.* Philadelphia, Pa: American Board of Internal Medicine, 1985.
14. Lewis M, Elwyn G, Wood F. Appraisal of family doctors: an evaluation study. *Br J Gen Pract* 2003; **53:** 454–460.
15. Beehr T, Ivanitskaya L, Hansen C. Evaluation of 360 degree feedback ratings: relationships with each other and with performance and selection predictors. *Journal of Organizational Behaviour* 2001; **22(7):** 775–788.
16. Church A. Do you see what I see? An exploration of congruence in ratings from multiple perspectives. *J Appl Soc Psychol* 1997; **27(11):** 983–1020.
17. Woolliscroft J, Howell J. Resident–patient interactions: the humanistic qualities of internal medicine residents assessed by patients, attending physicians, program supervisors and nurses. *Acad Med* 1994; **69:** 216–224.
18. Wenrich M, Carline J. Ratings of the performances of practicing internists by hospital-based registered nurses. *Acad Med* 1993; **68:** 680–687.
19. Davis J. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. *Obstet Gynecol* 2002; **99:** 647–651.
20. Van Zwanenberg T. Revalidation: the purpose needs to be clear. *BMJ* 2004; **328(7441):** 684–686.
21. Lakhani M. GMC and the future of revalidation: a way forward. *BMJ* 2005; **330:** 1326–1328.
22. Van den Bos K. Procedural and distributive justice: what is fair depends more on what comes first than on what comes next. *J Pers Soc Psychol* 1997; **72:** 95–104.