BJPsych The British Journal of Psychiatry (2008) 193, 156–160. doi: 10.1192/bjp.bp.107.041681

Questionnaires for 360-degree assessment of consultant psychiatrists: development and psychometric properties

Paul Lelliott, Richard Williams, Alex Mears, Manoharan Andiappan, Helen Owen, Paul Reading, Nick Coyle and Stephen Hunter

Background
Expert clinical judgement combines technical proficiency with humanistic qualities.

Aims
To test the psychometric properties of questionnaires to assess the humanistic qualities of working with colleagues and relating to patients using multisource feedback.

Method
Analysis of self-ratings by 347 consultant psychiatrists and ratings by 4422 colleagues and 6657 patients.

Results
Mean effectiveness as rated by self, colleagues and patients, was 4.6, 5.0 and 5.2 respectively (where 1=very low and 6=excellent). The instruments are internally consistent (Cronbach's alpha >0.95). Principal components analysis of the colleague questionnaire yielded seven factors that explain 70.2% of the variance and accord with the domain structure. Colleague and patient ratings correlate with one another (r=0.39, P<0.001) but not with the self-rating. Ratings from 13 colleagues and 25 patients are required to achieve a generalisability coefficient ($E\rho^2$) of 0.75.

Conclusions
Reliable 360-degree assessment of humane judgement is feasible for psychiatrists who work in large multiprofessional teams and who have large case-loads.

Declaration of interest
The Royal College of Psychiatrists manages the ACP 360 system and charges a fee for its use by consultant psychiatrists. Funding detailed in Acknowledgements.

Expert clinical judgement combines technical proficiency, which is derived from scientific understanding, with humanistic qualities. The latter are of particular importance in psychiatry where interprofessional teamwork and a good working relationship with the patient are both essential to high-quality care. Techniques to measure the technical competency of doctors are better developed than those to measure the non-technical aspects of clinical practice. One reason is that important aspects of technical proficiency are observable as behaviours and therefore are measurable using, for example, clinical audit and workplace-based assessment of competencies and skills. Our intention has been to develop and test an approach to assessing the qualities that constitute good humane judgement by senior psychiatrists (consultants). This paper describes the development of a 360-degree assessment system and the analysis of the results for the first 347 psychiatrists to participate.

Methods

Development of the questionnaires

Research ethics approval was obtained for all stages of the work to develop the questionnaires.

Theoretical background

Two of the seven core domains defined by the UK's General Medical Council (GMC) as being central to good medical practice encapsulate the humane qualities that are required by doctors.[1] The domain 'relationships with patients' includes good communication, being open and honest, and the quality of the doctor's relationships with the relatives, carers and partners of patients. The domain 'working with colleagues' includes effective teamworking, respect for colleagues and appropriate sharing of information.

We set out to express these desirable interpersonal qualities as observable behaviours and then to measure them using multisource feedback from the colleagues and patients affected by the consultant's work performance. This type of '360-degree assessment' has been applied to physicians and surgeons.[2–6]

The stated purpose of assessment may affect the ratings given by an appraiser.[7] We therefore informed all involved in developing and testing the questionnaires that their purpose was to give constructive feedback to the consultant as part of a developmental, 'formative' process with an emphasis on personal and professional development; for example, as one component of appraisal. It was not intended that the results be used as a summative judgement, for example about fitness to practice.

Identification of items and initial testing

We used the critical incident technique[8,9] in interviews with 24 specialist mental healthcare workers (four consultant and six non-consultant grade psychiatrists, five nurses, three other clinical staff and six managers). We asked interviewees to think of the last time they had seen a consultant do something that was particularly effective and something that was particularly ineffective in relating to patients or working with colleagues. For each event, the interviewee was asked: what was the situation; what events led up to it; what exactly happened; why was it effective/ineffective; and what was the outcome?

Two raters independently undertook a thematic analysis of interview transcripts and then met to generate the first index of behavioural competency items. This was presented to a focus group comprising four psychiatrists, two nurses, one other practitioner, three managers and two medical secretaries.

Participants ranked and weighted the behavioural competencies and used the repertory grid technique[10] to draw on their personal experience to relate the competencies to the expected performance of: an expert psychiatrist; a novice psychiatrist; an experienced psychiatrist but one who would not be considered an expert; the worst psychiatrist with whom they have worked; an expert, non-psychiatric doctor; and a trainee psychiatrist.

The project team edited the items to ensure that each described a different behaviour and that each was worded so that its meaning was unambiguous. The items were grouped into domains to create the first draft questionnaires – one for self-rating and one for completion by colleagues. Minor revisions were made following an initial test of face validity and feasibility by eight consultants and their colleagues. Finally, a subset of 17 items was selected for a questionnaire that could be completed by patients.

Piloting

Fifty-one consultant psychiatrists participated in the pilot. They had volunteered in response to a letter of invitation sent to members of the Faculty of General and Community Psychiatry of the Royal College of Psychiatrists. The 51 consultants, 609 of their colleagues and 937 of their patients completed questionnaires. The consultants and the colleagues were also asked to rate the importance of each item on a scale of 1=unimportant to 6=essential.

The mean rating of importance for the 59 items was 5.1 for the self-rating questionnaire and 5.3 for the colleague version. The lowest mean rating of importance for any individual item was 4.4; a score of 4 on the importance scale equates to 'important'. The inter-item correlation between two pairs of items in both the self-rating and colleague questionnaire was above 0.8. One item of each pair was dropped leaving the same 57 items in the final versions of the self-rating and colleague questionnaires.

The final questionnaires

The 57 items in the self-rating and colleague questionnaires cover nine domains:

(a) communication (six items)

(b) availability (four items)

(c) emotional intelligence (eight items)

(d) decision-making (seven items)

(e) relationships with patients (nine items)

(f) relationships with patients' relatives, partners and carers (five items)

(g) relationships with consultant colleagues (four items)

(h) relationships with junior doctors (six items)

(i) relationships with the team and external agencies (eight items).

Each item describes briefly a behaviour that relates to the domain in question. For example, items in the emotional intelligence domain include 'offering reassurance when appropriate', 'being consistently respectful of others' and 'being willing to take advice from others'.

The patient questionnaire contains 17 items. The behaviours that these describe have their counterparts in items in the first six domains of the self-rating and colleague questionnaires.

All three questionnaires require that each person who completes it rates the effectiveness of the subject for each item by scoring observed behaviours (from 1=very low to 6=excellent). The instructions for the colleague questionnaire state that the purpose of the assessment is so that the consultant can 'learn about themselves and develop'. They ask each person who completes the questionnaire to be honest, to ensure that one aspect of the consultant's personality does not influence all ratings, to use current behaviour as the basis for ratings and to be neither overly favourable nor overly critical. The patient questionnaire states that the purpose of assessment is so that the consultant 'can improve the quality of care he or she provides'. Both questionnaires state that a consultant will not be able to identify the rating made by any individual responder.

The three questionnaires that resulted from the process of development and initial testing then became the core of a new 360-degree assessment service for consultant psychiatrists (ACP 360) that is offered by the Royal College of Psychiatrists. The full questionnaires, and their accompanying instructions, can be viewed at www.rcpsych.ac.uk/crtu/centreforqualityimprovement/acp360.aspx.

Recruitment

The ACP 360 was launched by the Royal College of Psychiatrists in Autumn 2005. A recruitment letter was sent to all consultant psychiatrists who work with individuals of working age in the UK. There was a charge for participation. We present here the results of an analysis of the returns from the first 347 participants.

The assessment process

Each consultant was sent a pack, which included written instructions. They were asked to complete a rating questionnaire about themselves, either in paper form or online, and to select at least 15 colleagues and 30 patients to act as appraisers. It was recommended that the selected colleagues included one line manager (clinical/medical director or chief executive), four consultant psychiatrists, seven other clinical colleagues (such as team manager, nurses, social workers, occupational therapists, psychologists, junior doctors) and three non-clinical colleagues (such as secretaries, clerks or administrators). Consultants were asked to select patients with whom they had had significant and recent contact and for their selection to reflect their case-load in terms of gender, age, ethnicity and diagnosis.

Each consultant requested each of their colleagues to complete a questionnaire, either online or in paper form, and each consultant posted a paper questionnaire to the 30 patients together with a standard letter. Colleagues and patients sealed paper questionnaires in a pre-addressed envelope and returned them to the Royal College of Psychiatrists. The system is designed so that consultants do not know which particular persons have returned a questionnaire and they never see any completed questionnaires. The instructions explained this arrangement to colleague and patient raters. Ratings from a minimum of 10 colleagues and 10 patients were required before the returns were analysed. If this threshold had not been reached within a set period, the consultant concerned was asked to send a reminder letter to all 15 colleagues and 30 patients.

Data analysis

We examined questionnaires and ratings from three perspectives: self, colleague and patient, using descriptive statistics. We tested internal consistency using Cronbach's alpha. We used principal components analysis with varimax rotation, to study the structure of the colleague and patient questionnaires. We examined inter-rater reliability of both the colleague and patient questionnaires

https://doi.org/10.1192/bjp.bp.107.041681 Published online by Cambridge University Press

using intraclass correlations and used the generalisability coefficient,[11,12] denoted as $E\rho^2$, to estimate the number of colleague and patient ratings required to undertake a meaningful assessment of a consultant using this approach. Finally, we examined the relationships between the three measurement perspectives (self, colleague and patient) using Pearson correlation coefficients. All analyses were conducted using SPSS version 14.0.

Results

Response rates

A total of 4422 colleagues (a mean of 12.7 per consultant; range 10–17) and 6657 patients (a mean of 19.2 per consultant; range 10–31) rated the 347 consultants.

Missing data

A mean of 2.0% of data items were missing for the self-rating questionnaire; no single item was left unrated by more than 33 consultants (9.5% of the total). As regards the 57 items of the colleague questionnaire, 28 items were left unrated by fewer than 5% of colleagues, 20 by between 5% and 10%, and 9 by more than 10%. All 9 items with more than 10% of missing data were in the domains concerning relationships with consultant peers and with junior doctors. The percentage of missing data ranged from 10.4% to 21.1% for items in these two domains. Overall, 13 of the 17 items in the patient questionnaire had fewer than 5% missing data. The exceptions were 3 items relating to carers and family members (range of missing data, 12.8% to 19.4%) and an item asking whether the consultant remained calm under pressure (6.2% missing data). In subsequent analyses, missing data were replaced by mean imputation when necessary.

Ratings of effectiveness

The mean effectiveness ratings for all consultants for all items were 4.6 for the self-rating (s.d.=0.9), 5.0 for the colleague rating (s.d.=0.9) and 5.2 for the patient rating (s.d.=1.0). Overall, both colleagues (paired *t*-test: $t=12.1$, $P<0.001$) and patients (paired *t*-test: $t=18.0$, $P<0.001$) gave significantly higher ratings to the consultants than the consultants gave to themselves. Mean ratings of effectiveness were high (above 4.0) for all domains for both self- and colleague ratings (Table 1).

Internal consistency and structure of the questionnaires

Cronbach's alpha for the self-, colleague- and patient-rated questionnaires were 0.98, 0.98 and 0.97 respectively. It is considered that a coefficient above 0.8 indicates adequate internal consistency and reliability.

The principal components analysis, with varimax rotation, of the colleague ratings, yielded seven factors with an eigen value greater than 1. These seven factors accounted for 70.2% of the total variance in the data. These factors were highly consistent with the structure of the questionnaire, as defined by items having a factor loading greater than 0.4 (Table 2). Five of the factors contained all the items in the corresponding domain. The principal components analysis of the patient ratings yielded a single factor that explained 66.8% of the total variance. This factor included all 17 items.

Table 1 Mean domain scores for the self-ratings ($n=347$) and ratings by colleagues ($n=4422$)

Domain		Scores, mean (s.d.)	
		Self	Colleague
1	Communication	4.6 (0.8)	5.0 (0.9)
2	Availability	4.7 (0.9)	5.0 (0.9)
3	Emotional intelligence	4.6 (0.8)	4.9 (0.9)
4	Decision-making	4.7 (0.8)	5.0 (0.9)
5	Relationships with patients	4.7 (0.8)	5.1 (0.8)
6	Relationships with relatives and carers	4.5 (0.8)	5.0 (0.8)
7	Relationships with consultant peers	4.2 (0.9)	4.8 (0.9)
8	Relationship with junior doctors	4.7 (0.9)	5.0 (0.8)
9	Relationship with team and external agents	4.5 (0.9)	4.9 (0.9)

Interrater reliability and generalisability

The intraclass correlation coefficient for the colleague questionnaire was 0.75. Ratings from 10 colleagues are required for an $E\rho^2 \geqslant 0.70$; and from 13 colleagues for an $E\rho^2 \geqslant 0.75$. Overall, 198 consultants (57%) achieved returns from 13 or more colleagues. For the patient questionnaire, the intraclass correlation coefficient was 0.70. A total of 19 patient ratings are required for an $E\rho^2 \geqslant 0.70$ and 25 for an $E\rho^2 \geqslant 0.75$. The lower number of patient returns was achieved by 183 consultants (53%) and the higher by 57 consultants (16%).

Relationship between the different ratings

The global self-rating, expressed as a mean of all items, was not significantly correlated with either the global colleague rating ($r=0.06$, $P=0.29$) or the global patient rating ($r=0.01$, $P=0.82$). The correlation between the global colleague rating and the global patient rating was significant ($r=0.39$, $P<0.001$). Table 3 shows the correlations between the equivalent domain scores for self- and colleague ratings, and between the patient ratings and the domain scores for both the self- and colleague ratings.

Discussion

The properties of ACP 360 and its limitations

Consistent with multisource feedback tools developed for other specialty groups, ACP 360 ratings from all three groups of raters are skewed to the positive end of the scale and colleagues tend to rate consultants higher than consultants rate themselves.[5,6,13]

With ACP 360, the appraisees decide which colleagues and patients will assess them. It is possible that the high scores are at least in part because of the exclusion of colleagues and patients that consultants believe would rate them poorly. However, there is some evidence that colleague ratings obtained by multisource feedback are the same whether colleagues are selected at random to participate or whether the doctor concerned makes the selection.[2] Also, a consultant working in a mental health team would need to involve a substantial proportion of colleagues to achieve the required number of ratings. We could find no study that examined the effect of method of selection on the ratings made by patients. It would be difficult to exclude all selection bias for patient raters because it is the psychiatrist concerned who is best placed to decide which patients should be asked. For example, some patients may lack the capacity to complete a questionnaire or might be distressed by being asked.

A weakness of both the colleague and patient questionnaires is that a substantial number of colleagues and patients are unable to rate some items. In particular, some colleagues are unable to assess the consultant's relationship with other doctors. This might be because, in the UK, some psychiatrists work in partial isolation from other consultants. Also, colleagues from other disciplines

https://doi.org/10.1192/bjp.bp.107.041681 Published online by Cambridge University Press

Table 2 The seven factors derived from the principal components analysis of ratings by colleagues and how these map to the questionnaire domains

Factor (% variance explained)	Description of factor	Equivalent domain(s)[a]	Domain items in this factor/total items[b]
A (17.4)	Relationship with patients and carers	5 & 6	14/14
B (14.6)	Emotional intelligence	3	7/8
C (9.1)	Relationship with junior doctors	8	6/6
D (9.0)	Relationship with team and external agencies	9	8/8
E (8.6)	Communication and relationship with consultant peers	1 & 7	6/10
F (6.1)	Decision-making	4	7/7
G (5.4)	Availability	2	4/4

a. This refers to the nine domains in the colleague rating questionnaire.
b. This column shows the number of items in the questionnaire domain(s) that have a factor loading of greater than 0.4 for the factor concerned e.g. seven of the eight items in the questionnaire domain for emotional intelligence are in Factor B.

Table 3 Pearson correlation coefficients between the self- and colleague, self- and patient, and patient and colleague ratings

Domain	Correlation		
	Self- and colleague ratings for the equivalent domain	Domains of self-ratings and mean of all patient items	Domains of colleague ratings and mean of all patient items
1 Communication	0.05	0.01	0.36*
2 Availability	0.22*	−0.05	0.22*
3 Emotional intelligence	0.10	0.02	0.35*
4 Decision-making	0.06	0.00	0.32*
5 Relationships with patients	0.01	0.00	0.34*
6 Relationships with relatives and carers	0.03	0.01	0.42*
7 Relationships with consultant peers	0.14*	−0.02	0.28*
8 Relationship with junior doctors	0.13*	−0.02	0.22*
9 Relationship with team and external agents	0.12*	0.01	0.30*

*$P<0.05$.

might have little opportunity to observe the working relationship between consultants and junior doctors. Some patients are unable to rate items concerning the relationship between the consultant and the patient's carers and family members. Perhaps carers and family members should be included as a separate, fourth group of raters in the 360-degree assessment process.

Three independent sources provide evidence of face and content validity for the instruments. First, the items were selected to cover certain components of performance described by the GMC as central to good medical practice[1] and as extended for psychiatrists by the Royal College of Psychiatrists.[14] Second, the development process involved three phases of research and consultation that involved both psychiatrists and those colleagues who would be asked to rate the psychiatrists. Third, those people who participated in the second, full-scale pilot rated all items as important, in both the self- and the colleague questionnaires. The high rate of return from both colleagues and patients suggests that raters did not find the questionnaires over-burdensome; although the respondents were probably highly motivated because they or their service had paid for them to participate. The disaggregation of complex humane attributes, such as 'availability', into a number of items that describe actual, observable behaviours might have made the questionnaires easier to use.

The three questionnaires show good internal consistency. Also, the colleague questionnaire has a factor structure that adheres to the domains covered by the items. Furthermore, both the domains that emerged from the qualitative work of developing the questionnaires, and the factors that were derived from the principal components analysis of the colleague questionnaire, are meaningful and mutually supporting. The patient questionnaire has a different structure to the colleague questionnaire. The 17 items, which correspond to items drawn from six of the domains of colleague questionnaire domains, form a single factor.

As with other 360-degree assessment instruments, which have been criticised for their low interrater reliability,[15] the intraclass correlations were modest for both the colleague and patient questionnaires. It has been proposed that an Ep^2 of 0.75 is the minimum requirement for generalisability for instruments used in multisource feedback.[15] Using this benchmark, only 57% and 16% of participating consultants achieved sufficient returns for the colleague and patient questionnaires respectively. As a result of this analysis we have increased the target number of returns for consultants participating in ACP 360 to 13 colleague and 25 patient questionnaires. This number of colleague ratings can be achieved by consultants who, as in the UK, work as part of a large, multiprofessional team with a team of administrative support staff that work closely with the consultant. Some consultants participating in ACP 360 extend their colleague cohort by including general practitioners who take on shared care responsibilities. The ACP 360 would not be suitable for psychiatrists who work in isolation and/or have small case-loads of patients.

A recent systematic review of studies that compared physicians' self-rated assessments with those of external observers concluded that 'the preponderance of evidence suggests that physicians have a limited ability to self-assess'.[16] The lack of correlation between self-ratings using ACP 360 and those made by colleagues is consistent with this conclusion. The authors of the review argue that this feature of self-assessment justifies the introduction of multisource feedback, 'particularly when interpersonal skills, communication skills or professionalism need to be evaluated'.[16]

In common with other multisource feedback instruments, we have not formally tested the validity of the ACP 360

https://doi.org/10.1192/bjp.bp.107.041681 Published online by Cambridge University Press

questionnaires.[15] There is no gold standard test with which to compare the results and in the UK there is no effective system for grading the performance of doctors other than at the extremes where doctors are subject to disciplinary procedures by their employers or have sanctions imposed by the GMC. However, colleagues and patients are two independent groups of raters whose knowledge of consultants is derived from quite different perspectives. Arguably, the fact that their ratings are significantly correlated is a measure of concurrent validity.

The role of 360-degree assessment in measuring the performance of doctors

Consultants from all medical specialties, who work in the four UK National Health Services, are subject to annual appraisal of their roles, work and performance. In the UK, structured feedback on performance is also used in other ways. It is an important element of the annual review of each consultant's job plan that is part of their contract of employment. It will play an important role in the new systems being introduced for organising and assessing medical training at all levels[17,18] and will be a core part of the new procedures for revalidation.[19]

The Chief Medical Officer for England has stated that 'while patients want their doctors to have good clinical knowledge and technical skills, they also rate the interpersonal aspects of care as equally, if not more, important'.[20] Although critics have questioned whether humanistic qualities can be measured validly,[15,20] 360-degree assessment, which will be a requirement for relicensing by the GMC,[19] is the assessment approach most likely to give meaningful results.

Psychiatrists participate in the Physician Achievement Review (PAR) Program that is managed by the College of Physicians and Surgeons of Alberta in Canada[22] and is probably the longest established 360-degree assessment system for doctors. However, the PAR questionnaires that are applied to psychiatrists have been adapted from questionnaires designed for general physicians. The authors are not aware of any reports of 360-degree assessment systems developed specifically for psychiatrists or whose focus has been specifically on the humanistic aspects of psychiatric practice.

Implementation of ACP 360

The Royal College of Psychiatrists has introduced ACP 360 as a service for its members in the UK (www.rcpsych.ac.uk/crtu/centreforqualityimprovement/acp360.aspx). Participation is voluntary and the system is funded by a subscription fee that is paid either by the consultant or by their employing organisation. The patient questionnaire has been adapted for assessing psychiatrists who work with children and adolescents and for psychiatrists who work with people with an intellectual disability and with older people. We are also exploring the use of the system for assessing non-consultant grade psychiatrists.

The ACP 360 has been designed, tested and implemented for formative purposes. It is not intended as a summative review but to inform consultants about how they might focus their personal development plans in order to continue to develop their performance as a component of striving towards excellence. Therefore, its purpose is to give feedback to individual consultants and the results are provided confidentially to each consultant who takes part. However, to provide a comparator, the ratings for each participant are also presented in the context of a mean 'benchmark' rating of all consultants who have participated previously.

Paul Lelliott, MRCPsych, Royal College of Psychiatrists' Research and Training Unit, London; **Richard Williams**, TD, FRCPsych, University of Glamorgan and University of Central Lancashire and Gwent Healthcare NHS Trust, Welsh Institute for Health and Social Care, St Cadoc's Hospital, Newport, Wales; **Alex Mears**, PhD, **Manoharan Andiappan**, MSc, Royal College of Psychiatrists' Research and Training Unit, London; **Helen Owen**, **Paul Reading**, both formerly at the National Leadership and Innovation Agency for Healthcare, Llanharan, Wales; **Nick Coyle**, MSc, Royal College of Psychiatrists' Research and Training Unit, London; **Stephen Hunter**, FRCPsych, Gwent Healthcare NHS Trust, Cwmbran, UK

Correspondence: Paul Lelliott, Royal College of Psychiatrists' Research and Training Unit, Standon House, 21 Mansell Street, London E1 8AA, UK. Email: plelliott@cru.rcpsych.ac.uk

First received 18 Jun 2007, final revision 9 Dec 2007, accepted 14 Dec 2007

Acknowledgements

The Wales Office for Research and Development funded the qualitative development work. Gwent Healthcare NHS Trust funded the first pilot, which was undertaken by the Welsh Institute for Health and Social Care in the University of Glamorgan and the National Leadership and Innovation Agency for Healthcare in Wales. The Royal College of Psychiatrists funded the second, full-scale pilot that was undertaken by the Royal College of Psychiatrists' Research and Training Unit.

References

1 General Medical Council. *Good Medical Practice*. GMC, 2006.

2 Ramsey PG, Wenrich MD, Carline JD, Innu TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA* 1993; **269**: 1655–60.

3 Violato C, Marini A, Towes J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Acad Med* 1997; **72** (suppl 1): S82–4.

4 Hall W, Violato C, Lewkonia R. Assessment of physician performance in Alberta: the Physician Achievement Review. *CMAJ* 1999; **161**: 52–6.

5 Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. *BMJ* 2003; **326**: 546–8.

6 Archer JC, Norcini J, Davies A. Use of SPRAT for peer review of paediatricians in training. *BMJ* 2005; **308**: 1251–3.

7 Antonioni, D. Designing an effective 360-degree appraisal feedback process. *Organisational Dynamics* 1996; **25**: 24–38.

8 Flanagan JC. The critical incident technique. *Psychol Bull* 1954; **51**: 327–58.

9 Dunn WR, Hamilton DD. The critical incident technique – a brief guide. *Med Teach* 1986; **8**: 207–15.

10 Kelly GA. *The Psychology of Personal Constructs*. Norton, 1963.

11 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979; **86**: 420–8.

12 Crocker L, Algina J. *Introduction to Classical and Modern Test Theory*. Holt, Rinehart and Winston, 1986.

13 McDermott A, Hasler J. 360° feedback: how do perceptions of doctors' attributes compare? *Clin Governance Bull* 2004; **5**: 3–4.

14 Royal College of Psychiatrists. *Good Psychiatric Practice* (2nd edn) (Council Report CR125). Royal College of Psychiatrists, 2004.

15 Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. *BMJ* 2004; **328**: 1240.

16 Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA* 2006; **296**: 1094–102.

17 Davies H, Archer J, Heard S, Southgate L. Assessment tools for foundation programmes – a practical guide. *BMJ* 2005; **330**: 195–6.

18 National Health Service. *Modernising Medical Careers*. NHS, 2006 (http://www.mmc.nhs.uk/pages/assessment/msf).

19 Department of Health. *Trust, Assurance and Safety: The Regulation of Health Professionals*. Cm 7013. TSO (The Stationery Office), 2007.

20 Baker R. Commentary: can poorly performing doctors blame their assessment tools? *BMJ* 2005; **330**: 1254.

21 Donaldson L. *Good Doctors, Safer Patients*. Department of Health, 2006.

22 College of Physicians and Surgeons of Alberta. *Physician Achievement Review*. 2006. (http://www.par-program.org/PAR-Info.htm).

https://doi.org/10.1192/bjp.bp.107.041681 Published online by Cambridge University Press