

Learning in practice

Review of instruments for peer assessment of physicians

Richard Evans, Glyn Elwyn, Adrian Edwards

Primary Care
Group, Swansea
Clinical School,
University of Wales
Swansea, Swansea
SA2 8PP

Richard Evans
clinical research fellow
Glyn Elwyn
professor

Adrian Edwards
reader

Correspondence to:
R Evans
richevansg@
msn.com

BMJ 2004;328:1240-3

Abstract

Objectives To identify existing instruments for rating peers (professional colleagues) in medical practice and to evaluate them in terms of how they have been developed, their validity and reliability, and their appropriateness for use in clinical settings, including primary care.

Design Systematic literature review.

Data sources Electronic search techniques, snowball sampling, and correspondence with specialists.

Study selection The peer assessment instruments identified were evaluated in terms of how they were developed and to what extent, if relevant, their psychometric properties had been determined.

Results A search of six electronic databases identified 4566 possible articles. After appraisal of the abstracts and in depth assessment of 42 articles, three rating scales fulfilled the inclusion criteria and were fully appraised. The three instruments did not meet established standards of instrument development, as no reference was made to a theoretical framework and the published psychometric data omitted essential work on construct and criterion validity.

Rater training was absent, and guidance consisted of short written instructions. Two instruments were developed for a hospital setting in the United States and one for a primary care setting in Canada.

Conclusions The instruments developed to date for physicians to evaluate characteristics of colleagues need further assessment of validity before their widespread use is merited.

Introduction

It is no longer enough to do a job to the best of one's ability. Other people have to be assured that professionals can be trusted, and interest is growing in the concept that colleagues might be well placed to make these judgments. We live in a society in which we are held to account for our performance, especially if we perform professional functions, and doubly so if these are funded for the common good, as is the case in medicine and education.¹ Interest therefore exists in how to measure the performance of doctors and other healthcare professionals. The recent focus on appraisal systems, recertification, revalidation, and continuous professional development bears witness to the interest in how to assess the ability of clinicians to maintain and

sustain their competence and to exhibit the qualities deemed to be necessary in their professional role.


Imbalances in knowledge between lay people and professionals make it difficult for lay people to assess doctors' ability and competence. Thus the idea of asking peers to assess professional performance, particularly the humanistic non-cognitive aspects (for example, qualities such as integrity, compassion, and responsibility to others) that are less accessible to conventional means of assessment such as written and clinical examinations, has been increasingly explored in the literature.²⁻³ But doubts remain about the validity of peer ratings of these aspects,⁵⁻⁷ where "high reliability, together with greater ease of use, may distract from concerns about validity when considering peer ratings as a measure of actual quality."⁸

We aimed to identify all existing instruments for rating peers in medical practice and to evaluate them in terms of how they were developed, their validity and reliability, and their appropriateness for use in clinical practice, including the primary care setting.

Methods

We did a systematic search for references to instruments for the rating of physicians by peers in the world literature. The search strategy included keyword combinations—for example, physician review, peer evaluation, colleague assessment. We searched Medline, 1966 to present; Embase, 1980 to present; PsycINFO, 1972 to present; ASSIA for Health; CINAHL; and the Cochrane Database of Systematic Reviews (see bmj.com for more details).

We included instruments if they had been specifically developed for use by physicians for the review or assessment of a peer or colleague in practice. We required articles to have some data on either the way the instruments were developed or their validation using psychometric methods. We excluded instruments if they were designed primarily for self completion, those not completed by physician peers, and instruments that were designed for use in purely educational settings.

 Extra tables and details of excluded instruments are on bmj.com

 This is an abridged version; the full version is on bmj.com

We examined each instrument identified in terms of its purpose (explicit or implicit), whether it had a developmental aim (formative), or whether the assessment was intended to identify a standard or judgment (summative). We assessed the theoretical underpinning, if specified, and how the tool had been developed and evaluated. We compared the samples, including the ratio of peers to index physician, and the total number of questionnaires considered in the psychometric analyses. We examined the method of identifying peers, how anonymity (or otherwise) of ratings was managed, the existence of benchmarks, and whether instruction or training was provided for the raters.

Results

The search identified 4566 articles; we obtained full papers for all 42 potentially relevant articles. Eighteen papers related to instruments for the rating of physicians by peers. We finally included three instruments: the professional associate rating,^{9 10} the peer assessment questionnaire,^{2 11} and the peer review evaluation form.¹² All three were developed in North America or Canada.

Professional associate rating—This is a questionnaire from the United States that consists of a scale for rating fellow physicians on a range of parameters based on American Board of Internal Medicine recommendations and encompassing clinical competence, communication skills, and humanistic qualities. The board uses the professional associate rating as part of its continuous professional development programme.¹³ This programme has three components: self evaluation, a secure examination (single best answer questions), and verification of credentials. The self evaluation component includes an elective “patient and peer assessment module,” which includes the professional associate rating and also patient ratings, self ratings, and a quality improvement plan.

Peer assessment questionnaire—This Canadian instrument uses a rating scale covering the dimensions of clinical competency, professional management, humanistic communication, and psychosocial management.^{2 11} This was used with other instruments

to produce multisource assessment known as 360 degree feedback, including patients, coworkers (non-physicians), and self. The development of the questionnaire was based on a grid of competences derived from a professional committee, with further development through focus groups. This was reported as clarification of wording and deletion of inappropriate items, but assessment of construct validity was not reported.

Peer review evaluation form—This instrument from the United States consists of a scale for rating along dimensions derived from the American Board of Internal Medicine recommendations, including technical skills (obtaining history, examining, investigating) and interpersonal skills (demonstrating integrity, empathy, and compassion).¹² The authors noted that specific training in the use of such an instrument would require residents and faculty to mutually define terms such as integrity, empathy, and compassion. It is also worth noting that linking multiple assessment factors such as integrity, empathy, and compassion in a single item poses potential dilemmas for the rater.

In summary, some instruments seem to be described in the literature, but only three have psychometric data about either their development or their validity and reliability. The table shows the essential characteristics of these three instruments (shown in more detail in tables A and B on bmj.com). None of the identified instruments refers to a theoretical framework. Other than factor analysis performed on the empirical results, little other psychometric assessment has been undertaken, and an important omission is the lack of attention given to construct and criterion validity.¹⁴ Explicit purposes of the instruments are either unmentioned or evolve over time. For the professional associate rating, a generalisability coefficient of 0.7 is quoted, suggesting good levels of reliability, whereas standard psychometric texts recommend coefficients of 0.75 as “a fairly minimal requirement for a useful instrument.”¹⁴ Moreover, concentrating on reliability and feasibility is premature when concerns exist regarding validity. The developers of none of the three instruments examined had addressed how to guide or train the assessors, other than by providing written instructions.

Data from three instruments, comparing theoretical base, purpose, and comments

Country and author	Theoretical base	Purpose	Setting	Scale descriptions	Comments
USA Ramsey et al 1996 Professional associate rating ^{9 10 15}	Unspecified—to isolate and measure characteristics of “good physician,” reflects domains of ABIM recommendations for evaluating humanistic qualities	Shift from initially implied performance ³ —potential feedback to practitioner became explicitly formative ¹⁰	Hospital general internists (228)	9 point Likert-type scale; 11 items, four of which were respect, integrity, compassion, and responsibility; end descriptors—eg, “Does not accept responsibility . . . fully accepts responsibility”	Development pathway and validity unclear; reliability and feasibility confirmed in US hospital physicians—applicability to UK general practice setting uncertain; now adopted as part of “patient and peer assessment module” of ABIM’s CPD programme ¹⁰ as formative instrument
Canada Hall et al 1999 Peer assessment questionnaire ^{2 11 16}	Broad principles of 360 degree or multisource feedback	Primarily formative and CPD; explicit aim for quality improvement by education rather than identification of “bad apples”	Mix of mainly family physicians (251) and clinical specialists (57) (35% rural)	5 point Likert-type scale; 24 items: 1=among the worst, 2=bottom half, 3=average, 4=top half, 5=among the best	Development path, including focus group, described; subject sample included family practitioners; part of 360 degree multisource feedback approach; for PAQ, clinical competency factor dominant (73%); for PS and CAQ, humanistic-communication factor dominant (61% and 79.6%)
USA Thomas et al 1999 Peer review evaluation form ¹²	Constructed to reflect ABIM domains	Primarily formative and CPD: 1. to provide feedback; 2. to enhance skills of self assessment and feedback	Hospital internal residency training (16)	9 point Likert-type scale: superior (far exceeds expectations), satisfactory (meets expectations), unsatisfactory (falls short)	Acknowledged by authors that unknown criteria used to rate—may vary between different rater groups—and need for specific training in evaluation will require mutually defined meanings of terms such as integrity

ABIM=American Board of Internal Medicine; CAQ=co-worker assessment questionnaire; CPD=continuous professional development; PAQ=peer assessment questionnaire; PS=patient survey.

What is already known on this topic

The range of professional competences and qualities now recognised as necessary in a good physician is not adequately assessed by conventional examinations and assessments

Suitable methods are needed to assess the broader range of competences, including “humanistic” qualities and professionalism

Peers are one potential source of assessment of these aspects of physicians’ practice

What this study adds

Very few instruments designed for peer assessment of physicians exist, and their development so far has focused on reliability and feasibility

The available instruments lack theoretical frameworks, and their validity remains questionable

Clarity of purpose is a key determinant of the subsequent “success” of peer appraisal but may be lost by confounding summative and formative aims

Discussion

Considerable interest exists in the concept that physicians can assess each other across a range of qualities (for example, integrity, compassion, respect, and responsibility), but this review shows that the instruments developed for peer assessment have not been developed in accordance with best practice. The principles of instrument design involve giving attention to theoretical frameworks and construct clarification in order to establish validity as the basis for reliability studies. These steps are not described for the instruments we identified.

Caution is needed when developing quantitative measures that use peers to rate complex humanistic qualities, and the complex nature of this field should be acknowledged.¹⁷ A common theme in the assessment literature is the question of self evaluation versus external evaluation and whether “others” can form judgments on differing facets of professional practice.¹⁸ “Social comparison theory” acknowledges the drive to self evaluate, using similar others as a benchmark,^{19–21} and recognises the construct of “managerial self awareness” as a process of self reflection using feedback,²² allowing us to “see as others see.”¹⁸ The validity of “others” as appropriate rater groups remains a challenge for research, because criteria and frames of reference, even if defined explicitly, will vary with each individual.^{6 23 24} In other words, how many “true” peers do professionals have? How many peer colleagues are in positions of having accurate knowledge about an individual’s performance in terms of compassion, responsibility, or respect, so that they can make informed judgments?

The other key issue is the perceived fairness of the peer appraisal process. Procedural justice theory suggests that people naturally make judgments on how decisions are arrived at (procedural justice) quite separately from judgments on outcomes of decisions (distributive justice).²⁵ Procedural justice is seen to be more important than outcome in terms of overall acceptability and an essential element of validity. This initial judgment on fairness also sets a frame of reference for interpreting subsequent events that has a

crucial and enduring influence.²⁵ Doubts about the face validity of arriving at a judgment on a peer’s compassion or integrity risk jeopardising the peer appraisal process through negative perceptions, which could be difficult to overcome subsequently. Evidence seems to exist that an appraisal process, once underway, enters a feedback loop of success that quickly becomes positive or negative with no safe middle ground.¹⁷ The identified instruments would need to consider procedural justice by demonstrating their validity through clearly defined criteria and constructs relevant to the rater groups.

Concern has been voiced about the validity of peer evaluation. If the validity of peer ratings remains unclear, then reliability and feasibility are no substitute.^{3 8 9 15} A possible approach has been initiated in Finland, where qualitative methods have been used to begin to characterise some of the concepts and constructs relevant to peer appraisal that are needed before quantitative tools are developed.²⁶ When peers have attempted to rate humanistic qualities, the validity has not been well supported by empirical findings. The poor agreement between observers of the same events is shown by several studies.^{5–8} An argument is emerging that the most valid source of ratings for humanistic dimensions are patients,^{5 6 22} because only they have experienced certain qualities, such as “a level of intimacy,” not available to other raters such as peers.²²

Implications for policy

The validity of rating items in assessing aspects such as the compassion, integrity, respect, or responsibility of a peer remains highly suspect. To have any validity or reliability, such qualities would need to be expressed as observable behaviours. In the absence of clearly defined constructs derived from a bottom-up empirical approach, and lacking a coherent theoretical framework, what is being measured here, if anything, is unclear.

Implications for research

Interest exists in using peers to assess the humanistic qualities of physicians, but the theoretical underpinning is lacking. Clarity of purpose is vital, and more attention needs to be given to the underlying constructs of interest. That judgments can be made only by those people who experience the qualities in question must be recognised. In the meantime, peer assessment methods should be used with caution.

We acknowledge the support of the Department of Postgraduate Education for General Practice, University of Wales Cardiff.

Contributors: See bmj.com

Funding: None.

Competing interests: None declared.

- O’Neill O. *Reith lectures: a question of trust*. BBC, 2002 (www.bbc.co.uk/radio4/reith2002).
- Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Acad Med* 1997;72(suppl):82-4.
- Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA* 1993;269:1655-60.
- Ramsey P, Wenrich M. Peer ratings: an assessment tool whose time has come. *J Gen Intern Med* 1999;14:581-2.
- McLoed P, Tamlyn R. Faculty ratings of resident humanism predict patient satisfaction ratings in ambulatory medical clinics. *J Gen Intern Med* 1994;9:321-6.
- Wooliscroft 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;69:216-24.
- Goldman R. The reliability of peer assessments: a meta-analysis. *Eval Health Prof* 1994;17:3-21.

- 8 Saturno P, Heather Palmer R. Physician attitudes, self-estimated performance and actual compliance with locally peer-defined quality evaluation criteria. *Int J Qual Healthcare* 1999;11:487-96.
- 9 Ramsey P, Carline JD, Blank LL, Wenrich MD. Feasibility of hospital-based use of peer ratings to evaluate the performance of practicing physicians. *Acad Med* 1996;71:364-70.
- 10 Lipner R, Blank L. The value of patient and peer ratings in recertification. *Acad Med* 2002;77:S64-6.
- 11 Hall W, Violato C, Lewkonia R, Lockyer J, Fidler H, Toews J, et al. Assessment of physician performance in Alberta: the physician achievement review. *CMAJ* 1999;161:52-6.
- 12 Thomas P, Gebo K, Hellmann D. A pilot study of peer review in residency training. *J Gen Intern Med* 1999;14:551-4.
- 13 American Board of Internal Medicine. *Recertification: program for continuous professional development*. www.abim.org/cpd/cpdhome/index.htm (accessed 23 Apr 2004).
- 14 Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford University Press, 1995.
- 15 Ramsey P, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. Predictive validity of certification by the American Board of Internal Medicine. *Ann Intern Med* 1989;110:719-26.
- 16 Fidler H, Lockyer JM, Toews J, Violato C. Changing physicians' practices: the effect of individual feedback. *Acad Med* 1999;74:702-14.
- 17 Peirperl M. Conditions for the success of peer evaluation. *Int J Hum Resour Manage* 1999;10:429-58.
- 18 Atwater L. Self-other agreement: does it really matter? *Persomel Psychology* 1998;51:577-98.
- 19 Festinger L. A theory of social comparison processes. *Hum Relat* 1954;7:117-40.
- 20 Fox S, Ben-Nahum Z. Perceived similarity and accuracy of peer ratings. *J Applied Psychol* 1989;74:781-6.
- 21 Mumford M. Social comparison theory and the evaluation of peer evaluation: a review and some applied implications. *Persomel Psychology* 1983;36:867-81.
- 22 Church A. Do you see what I see? An exploration of congruence in ratings from multiple perspectives. *J Applied Soc Psychol* 1997;27:983-1020.
- 23 Kaplan C, Centor R. The use of nurses to evaluate houseofficers' humanistic behaviour. *J Gen Intern Med* 1990;5:410-4.
- 24 Davis J. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. *Obstet Gynecol* 2002;99:647-51.
- 25 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.
- 26 Vuorinen R, Tarkka M. Peer evaluation in nurses' professional development: a pilot study to investigate the issues. *J Clin Nurs* 2000;9:273-81.

(Accepted 30 January 2004)

Commentary: "Soft" assessment—an oxymoron?

Murray Lough

O wad some Power the giftie gie us,
To see oursels as ithers see us!

R Burns, 1798

The desire to assess the "softer" parts of practice is strong. In addressing this ambition, many of the tensions and controversies surrounding the assessment process are thrown into sharp focus. Exactly how one defines soft practice for the jobs carried out by, say, general practitioners or orthopaedic surgeons is open to wild speculation. One key message of the paper by Evans et al is the importance of those being assessed having a stake in their assessment; this may be the "missing link" that could make the assessment process less threatening.

Assessment has to be rigorous, not least because it should be fair. The subtleties of language highlight the difficulties—peer assessment and peer review have different connotations. In an era of individual accountability and clinical governance, many might question whether soft practice can ever be assessed with rigour.

Evans and colleagues in their paper concentrate on what could currently be considered the psychometric gold standard attributes of assessment.¹ Van der Vleuten would argue that acceptability and cost effectiveness are also crucial in laying down a system where busy people are assessing busy people.² No matter how we cut the cake, assessment is expensive in terms of both time and money.

It is interesting that the three papers identified by Evans and colleagues emanate from North America, where a culture of a high level of psychometric skills is possibly encouraged by a litigation conscious community. This may highlight the lack of such skills in other countries and could serve as a warning that, given the high stakes involved in such assessments, the result is either instruments of poor quality, as suggested by the paper, or no instruments at all.

In response to this it may be that in considering assessment of a softer side of practice we need to consider some other measures of validity.³ One example is consequential validity—that is, taking into consideration the consequences on those being assessed as part of the process.⁴ In a world dominated by numbers and

the desire to measure them, a wider debate about the strengths and limitations of formal psychometrics when considering the less tangible aspects of practice is urgently needed if a more valid assessment of doctors is to be achieved.

In considering how one might address the assessment of peers, particularly with reference to the complexities of the jobs done, one example is to consider assessments over time—that is, trends such as the longitudinal evaluation of performance (LEP) used for the assessment of dental trainees in Scotland.⁵ It covers a range of competencies on a nine point scale and concentrates on feedback, a crucial element in consequential validity.

Three hundred and sixty degree or multisource feedback provides judgments on strengths and areas for potential improvement as perceived by work colleagues both above and below in the hierarchy. Many see this very practical method as opening new opportunities in exploring humanistic qualities that are not easily assessed by more traditional methods.⁶ Predictably, however, there are limitations, not least with validity, but these are not insurmountable.

Perhaps more qualitative work on defining soft practice is required before designing any more instruments. Themes on areas of soft practice could be collated drawn from groups of doctors with additional input from patients to form the basis for more formal psychometric development and testing. Without this, soft assessment will be seen as a soft option and not given the place it deserves.

Competing interest: None declared.

- 1 Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. *BMJ* 2004;328:1240-3.
- 2 Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ* 1996;1:41-67.
- 3 Messick S. The psychology of educational measurement. *J Educ Measur* 1984;21:215-37.
- 4 Moss PA. Shifting conceptions of validity in educational measurement: implications for performance assessment. *Rev Educ Res* 1992;62:229-58.
- 5 Prescott LE, Norcini JJ, McKinlay P, Rennie JS. Facing the challenges of competency-based assessment of postgraduate dental training, longitudinal evaluation of performance (LEP). *Med Educ* 2002;36:92-7.
- 6 King J. 360° appraisal. *BMJ* 2002;324:S195-6.

NHS Education for
Scotland, Glasgow
G3 8BW

Murray Lough
assistant director

murray.lough@
nes.scot.nhs.uk