Peer assessment of competence

John J Norcini

Objective This instalment in the series on professional assessment summarises how peers are used in the evaluation process and whether their judgements are reliable and valid.

Method The nature of the judgements peers can make, the aspects of competence they can assess and the factors limiting the quality of the results are described with reference to the literature. The steps in implementation are also provided.

Results Peers are asked to make judgements about structured tasks or to provide their global impressions of colleagues. Judgements are gathered on whether certain actions were performed, the quality of those actions and/or their suitability for a particular purpose. Peers are used to assess virtually all aspects of professional competence, including technical and non-technical aspects of proficiency. Factors influencing the quality of those assessments are reliability, relationships, stakes and equivalence.

Conclusion Given the broad range of ways peer evaluators can be used and the sizeable number of competencies they can be asked to judge, generalisations are difficult to derive and this form of assessment can be good or bad depending on how it is carried out.

Keywords education, medical/*methods; peer review/*methods; *professional competence; educational measurement; reproducibility of results.

Medical Education 2003;37:539–543

Introduction

The act of making judgements on the performance of one's peers is ubiquitous and has formed the basis of the referral process in medicine and other professions for centuries. Throughout this long history, peers have been deployed in a variety of different ways to make judgements on the competence of their colleagues. However, the systematic study of this form of measurement is surprisingly new, with one of the first reviews of the literature in the area published by Topping in 1998.[1]

For the purposes of this paper, peers are considered to be doctors or doctors-in-training who are similar in level of education, specialisation and practice. Assessment implies that they are making judgements about the quality of a colleague's performance in the domain of competence related to patient care. Given the broad range of ways in which peer evaluators can be used and the sizeable number of competencies they can be asked to judge, it is not surprising that generalisations are difficult to derive and that this form of assessment can be good or bad depending on how it is carried out.

This instalment in the series on professional assessment attempts to summarise how peers are used in the evaluation process and whether their judgements are reliable and valid. Specifically, it describes the nature of the judgements peers can make, the aspects of competence they can assess, the factors influencing the quality of the results and the steps in implementation.

Nature of the judgements

Judgements about structured tasks versus global impressions

Peers are asked to make judgements about structured tasks or to provide their global impressions of colleagues. Heylings and Stefani provide an example of judging structured tasks in a large anatomy class.[2] First year medical students were asked to mark the clinical case studies of their peers. A current example with practising doctors involves the General Medical Council

Foundation for Advancement of International Medical Education and Research (FAIMER®), Philadelphia, Pennsylvania, USA

Correspondence: John J Norcini PhD, Foundation for Advancement of International Medical Education and Research (FAIMER®), 3624 Market Street, 4th Floor, Philadelphia, Pennsylvania 19104, USA. Tel.: 00 1 215 823 2170; Fax: 00 1 215 386 2321; E-mail: jnorcini@ecfmg.org

Key learning points

Peers are asked to make judgements about structured tasks or to provide their global impressions of colleagues. These judgements take the form of whether certain actions were performed, the quality of those actions and/or their suitability for a particular purpose.

The factors influencing the quality of peer assessments include reliability, relationships, stakes and equivalence.

Peers are used to assess virtually all aspects of professional competence and the quality of their assessments relates to the nature of the judgements and how they are used.

(GMC) programme to assess doctors who may be seriously deficient.[3] Three trained assessors (2 of whom are specialty-matched physicians) review the doctor's medical records, discuss selected cases, observe a series of consultations or other relevant activities, tour the workplace, interview colleagues and conduct a structured interview. For each of these activities, they make judgements about the quality of the specific performance they review or observe. A similar programme has been running since 1980 in Ontario, Canada.[4]

It is much more common to ask peers to provide their global impressions of the competence of the doctor being assessed.[5] In these instances, they are asked to consider the performance of a colleague over some period of time and to make judgements about his or her ability in one or more dimensions of competence. For example, Hefler developed a peer evaluation of the doctor–patient relationship, responsibility as a doctor, emotional stability, and overall competence.[6] Linn *et al.* developed a rating scale for 3rd year medical students, asking them to rate their peers along the dimensions of knowledge and relationship skills.[7] Similarly, Ramsey *et al.* developed a rating form that was sent to the colleagues of practising internists asking for an assessment of the internists' competence in a variety of different areas.[8]

These 2 types of judgements have both strengths and weaknesses. To produce reliable results, any assessment of competence must be based on an evaluation of several different encounters with patients because physician performance is case-specific.[9] Consequently, global impressions, which take account of a variety of different encounters over time, have an advantage over judgements made on structured tasks. However, global assessments can be rendered even if the evaluator has not directly observed the behaviour in question. For example, it is widely known that residents are not often observed taking a history or conducting a physical examination, but they receive ratings of these aspects of competence from their supervisors anyway.[10] In addition, global judgements are often influenced by the evaluator's general impression of the person being assessed rather than by his or her skill in a specific area. This halo effect can be reduced to some extent when a structured task is being evaluated.

Judgements about occurrence, quality or suitability

Peers can be asked to make judgements about whether a colleague performed certain actions, the quality of those actions and/or whether they were suitable for a particular purpose. In terms of making judgements about occurrences, peers can be asked to determine whether a colleague performed specific actions (i.e. filled out a checklist). For example, Calhoun *et al.* required all 2nd year students to be videotaped while performing a physical examination.[11] Peers then assessed the tapes by completing a checklist (e.g. auscultates the right lower quadrant, percusses the left lower quadrant). They were generally accurate although they tended to be more demanding and more similar to faculty when asked to repeat the same task with the same tapes 2 years later.

Most often, peers are asked to make judgements about the quality of their colleagues' performances. For example, Van Rosendaal and Jennett asked residents to rate the ability of their peers' to perform a physical examination.[12] To assure that the ratings derived from these and similar scales are meaningful, it is essential that peers have observed 1 or more performances and are able to make judgements about their quality. For trainees early in the educational process, it may not be reasonable to assume that they can make distinctions about the quality of performances. In contrast, practising doctors should be relatively good judges of quality but they may not have observed many of the behaviours they are asked to rate. The meaning of peer assessment will relate directly to how well these assumptions are met.

Occasionally, peers are asked to make 2 decisions at the same time: firstly in the form of a judgement about the quality of a performance and secondly relating to whether it is good enough for a particular purpose. For example, a peer rating form with anchors for 'satisfactory' and 'unsatisfactory' requires the evaluator to make a judgement about how good the performance was and then to decide whether it was good enough (i.e. 'satisfactory') for the purpose of the evaluation. There are errors associated with both of these forms of judgement and asking that they be combined and

completed by a single evaluator renders their meaning unclear. For example, if a performance is rated as 'unsatisfactory' by an evaluator, the reason may be that it was a poor performance or it may be that the evaluator has high standards. Where possible, separating these 2 judgements will improve the reliability and validity of the ratings.

Aspects of competence assessed

Peers are used to assess virtually all aspects of professional competence. For example, Risucci *et al.* asked residents to judge their peers' technical ability, basic science knowledge, clinical knowledge, judgement, peer relations, patient relations, reliability, industry, personal appearance and reaction to pressure.[13] Similarly Ramsey *et al.* developed a rating form that asked questions about practising doctors' ambulatory care skills, management of complex problems, management of hospitalised patients, problem-solving, integrity, sensitivity to psychosocial aspects of illness, compassion, responsibility and overall competence.[8]

Although peers were asked to make judgements across this broad range of competencies, they were not necessarily able to identify differences among all of them. Many of the ratings were highly correlated and from an analysis of the data for both residents and practising physicians, 2 factors emerged: technical or cognitive skills and non-cognitive or relationship skills.[7,8,14,15] It may be that differences among students/doctors within these 2 broad domains exist only rarely or that they exist often but are not discerned by peers.

The fact that peers do not make distinctions within the technical/cognitive and relationship/non-cognitive domains does not necessarily reflect poorly on the validity of peer assessment. To the contrary, several studies indicate that this form of evaluation has reasonable relationships with other measures of ability. Ratings of students and residents were correlated with grades given by faculty and written examination performance.[7,12,13,15–17] Likewise, ratings of practising physicians had a positive relationship with certification status and previous certifying examination scores.[8] These and other studies led Eva to conclude that peer assessment provides a valid form of tutorial-based evaluation (compared to faculty and self-assessment) and a similar conclusion was reached by Topping for peer assessment outside of medicine.[1,18]

Factors influencing the quality of peer assessment

Although peer assessment has the potential to provide accurate and valid assessment information, several factors will influence the quality of the results. Specifically, they are reliability, relationships, stakes and equivalence.

Reliability

Three major factors will contribute to the reliability of peer assessments: the number of relevant performances observed, the number of peers involved, and the number of aspects of competence being evaluated. There is a sizeable body of evidence indicating that doctor performance varies from patient to patient, so reliable results require information from a number of encounters.[9] For peer assessment, this means that evaluation should be based on observation in a variety of different clinical situations.

Similarly, the literature shows clearly that even experienced evaluators differ when observing exactly the same events.[19] Consequently, evaluations from several colleagues are needed to achieve reliable results. For example, Ramsey *et al.* estimated that 11 peers were required to achieve a reliability coefficient of 0.70.[8] In most measurement situations, increasing the number of evaluators will have a smaller impact on the reliability of ratings than increasing the number of encounters observed. However, in peer assessment this is less likely to be the case because additional peers often bring with them different observations.

Finally, it is important to ask for evaluations of several aspects of the competence being assessed. For example, Ramsey *et al.* found that reasonably reliable estimates of overall competence could be achieved with roughly 10 questions about different aspects of competence.[8] This number should be more than sufficient for most purposes, and has the advantage of making the scale long enough to obtain reliable results but not so long that it becomes burdensome and decreases the response rate.

There are a number of factors that have relatively small influences on reliability. Included in this category are issues such as the wording of questions, the number of points on rating scales, whether to describe all points on the scale, and the like. Unfortunately, because they are visible and easily manipulated, many users of peer assessment spend inordinate amounts of time addressing them. Assuming the application of minimal care, time is better spent recruiting and training peers.

Relationships

Although the interaction among peers provides a very rich source of assessment information, the nature of the relationships between the doctors being evaluated and their colleagues can pose difficulties. Students who compete with each other or who are personal friends

may be motivated in grading by more than relevant performance. Likewise doctors in practice might have a series of financial relationships that militate against the perception, if not the reality, of accurate and valid assessment. Consistent with the power of these interpersonal relationships, Van Rosendaal and Jennett reported that internal medicine residents felt strongly that peer assessment was an unwanted intrusion.[20]

Although there is no way to avoid this issue, 1 way of reducing its effect is to ensure that the evaluations remain anonymous. Under this restriction, Ramsey *et al.* collected ratings from 2 lists of peers, one provided by the doctor being assessed and the other provided by his or her chief of service.[8] They found no significant differences between the ratings supplied by these different groups, nor were there differences in the professional and social relationships they had with the doctor being evaluated.

Stakes

There is no definitive research on the influence of the stakes of an assessment on the validity of the results. However, it is reasonable to think that ratings, grades or feedback could be influenced by the use to which they will be put. To this point, Hay suggested that when peer assessment is used in a high stakes setting, it results in inflated estimates of performance and few below average evaluations.[21] In fact, Ramsey *et al.* and Hall *et al.* found that peers provided uniformly high ratings (better than 7 on a 9-point scale in the Ramsey study), with few doctors receiving ratings near the bottom of the scale.[22,23] Of course, it is unclear whether this is a reflection of the uniformly high quality of the study participants or of reluctance on the part of the peers to provide low ratings. It is, however, suggestive of Hay's contention that there is an influence.

Again, there is no way to completely avoid this issue but it may help to ensure the anonymity of evaluators. Further, it may be useful to limit the judgements to occurrences or the quality of performance rather than suitability for a particular purpose. By its nature, the latter is of higher stakes and setting standards for performance is better done in other ways.

Equivalence

One of the fundamental issues in peer assessment is whether the evaluation of a doctor or student is equivalent to that of his or her colleagues. Threats to equivalence come from 2 sources. Firstly, the activities doctors or students undertake while they are being judged may be different and therefore not of the same complexity. The growing use of portfolios recognises, and in some instances exacerbates, these naturally occurring differences in training and practice. Secondly, the group of peers may not be the same for all of the doctors being evaluated, and thus they may differ in stringency. In both instances, it is problematic to directly compare the assessments of students or doctors because the playing field is not level.

This is less of a problem at the level of the classroom or tutorial group, as all or most of the evaluators will be the same and control can be exercised over the performances being judged. For regional or national programmes, however, this is a bigger concern, as many doctors will be judged by completely different groups of peers based on their performance in a unique practice setting. These problems cannot be eliminated but increasing the number of peers involved in the process and providing them with clear criteria for making their judgements can minimise them.

Steps in implementation

There are at least 5 steps involved in implementing a process for peer assessment.[1] Firstly, the purpose of the assessment should be stated, preferably in writing. This purpose must be communicated to all participants, along with the expectations of their performance as both evaluators and as the objects of evaluation. Depending on the culture of the group or institution, it might be best to introduce this form of assessment gradually, perhaps starting with anonymous evaluations in a low stakes setting.

Secondly, assessment criteria must be developed and communicated to the participants. This includes how many and which peers will participate, what they will assess, when they will assess it, what constitutes the acceptable range of quality, and, if necessary, what is considered a suitable performance. Out of these criteria will flow the method(s) of data capture (e.g. checklist, rating form, scoring key) and the details for actually carrying out the assessments.

Thirdly, training should be provided to all the participants. This can range from simple written or verbal descriptions of what is expected to intense videotape-based benchmarking with feedback. In general, increasing the number of peers assessing each student will have a larger positive impact than increasing the intensity of training. If the number of peers is limited, however, more extensive training is probably needed.

Fourthly, the results of the assessments should be monitored throughout the implementation process. Simple checks on reliability and validity should be ongoing and feedback should be elicited from the participants. Additional training and corrections to data collection strategies can be carried out as needed.

Fifthly, feedback should be provided to participants. In their roles as evaluators, students should be compared to each other and those who are too stringent or too lenient should receive remediation. In their roles as the objects of evaluation, students should be given feedback appropriate to the purpose of the assessment. The entire process should be followed over time to assure that it is fulfilling its purpose.

Acknowledgements

This work was supported by the Foundation for International Medical Education and Research but does not necessarily reflect its opinions.

References

1 Topping K. Peer assessment between students in colleges and universities. *Rev Educational Res* 1998;**68**:249–76.
2 Heylings DJ, Stefani LA. Peer assessment feedback marking in a large medical anatomy class. *Med Educ* 1997;**31**:281–6.
3 Southgate L, Cox J, David T *et al.* The assessment of poorly performing doctors: the development of the assessment programmes for the General Medical Council's Performance Procedures. *Med Educ* 2001;**35**:2–8.
4 Norton PG, Dunn EV, Soberman L. What factors affect quality of care? Using the Peer Assessment Program in Ontario family practices. *Can Fam Physician* 1997;**43**:1739–44.
5 Arnold L, Willoughby L, Calkins V, Gammon L, Eberhart G. Use of peer evaluation in the assessment of medical students. *J Med Educ* 1981;**56**:35–42.
6 Helfer R. Peer evaluation: its potential usefulness in medical education. *Br J Med Educ* 1972;**6**:224–31.
7 Linn BS, Arostegui M, Zeppa R. Performance rating scale for peer and self assessment. *Br J Med Educ* 1975;**9**:98–101.
8 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.
9 Elstein AS, Shulman LS, Sprafka SA. *Medical Problem-Solving: an Analysis of Clinical Reasoning*. Cambridge, Massachusetts: Harvard University Press 1978.
10 Day SC, Grosso LG, Norcini JJ, Blank LL, Swanson DB, Horne MH. Residents' perceptions of evaluation procedures used by their training programme. *J General Intern Med* 1990;**5**:421–6.
11 Calhoun JG, Ten Haken JD, Woolliscroft JO. Medical students' development of self- and peer-assessment skills: a longitudinal study. *Teach Learn Med* 1990;**2**:25–9.
12 Van Rosendaal GM, Jennett PA. Comparing peer and faculty evaluations in an internal medicine residency. *Acad Med* 1994;**69**:299–303.
13 Risucci DA, Tortolani AJ, Ward RJ. Ratings of surgical residents by self, supervisors and peers. *Surg Gynecol Obstet* 1989;**169**:519–26.
14 DiMatteo MR, DiNicola DD. Sources of assessment of physician performance: a study of comparative reliability and patterns of intercorrelation. *Med Care* 1981;**19**:829–42.
15 Thomas PA, Gebo KA, Hellmann DB. A pilot study of peer review in residency training. *J Gen Intern Med* 1999;**14**:551–4.
16 Korman M, Stubblefield RL. Medical school evaluation and internship performance. *J Med Educ* 1971;**46**:670–3.
17 Schwarz RW, Donnelly MB, Sloan DA, Young B. Knowledge gain in a problem-based surgery clerkship. *Acad Med* 1994;**69**:148–51.
18 Eva KW. Assessing tutorial-based assessment. *Adv H Sci Educ* 2001;**6**:243–57.
19 Noel GL, Herbers JE, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med* 1992;**117**:757–65.
20 Van Rosendaal GM, Jennett PA. Resistance to peer evaluation in an internal medicine residency. *Acad Med* 1992;**67**:63.
21 Hay JA. Tutorial reports and ratings. In: Shannon S, Nocterm G (Eds). *Evaluation Methods: A Resource Handbook*. Hamilton, Ontario: McMaster University 1995.
22 Ramsey PG, Carline JD, Blank LL, Wenrich MD. Feasibility of hospital-based use of peer ratings to evaluate the performance of practising physicians. *Acad Med* 1996;**71**:364–70.
23 Hall W, Violato C, Lewkonia R, Lockyer J, Fidler H, Toews J, Jennett P, Donoff M, Moores D. Assessment of physician performance in Alberta the physician achievement review. *CMAJ* 1999;**161**:52–7.

Received 2 *October* 2002; *editorial comments to author* 10 *October* 2002; *accepted for publication* 13 *January* 2003