EDITORIALS

Peer Ratings
An Assessment Tool Whose Time Has Come

Peer ratings—ratings of physicians' clinical skills, communication skills, and professionalism completed by peer physicians or other professional colleagues—are coming into their own in many health care settings. For example, the American Board of Internal Medicine (ABIM) recently established professional associate ratings as a component of its future recertification programs.[1] Peer ratings are also being used or considered for evaluation of physicians in other settings, including academic departments, hospitals, HMOs, and large group practices. Although peer ratings are somewhat cumbersome when completed on paper because multiple ratings are required to achieve reliability, they remain one of the easiest and most comprehensive methods available to assess the skills of an individual physician. No other existing methodology provides practice-based information about physician performance in cognitive and interpersonal areas as well as professionalism. In upcoming years, use of computer and telephone technologies should reduce the time and expense required for collecting peer ratings and facilitate analysis of the information obtained. For example, the ABIM will use interactive voice response technology to obtain ratings by telephone rather than by mail.

In the medical school setting, ratings are frequently used, in concert with other assessment measures, to assess the performance of students on clinical rotations. Ratings by faculty and residents provide information about more domains of performance than a written examination alone can provide. Carline and colleagues studied the psychometric characteristics of ratings in a 12-week internal medicine clerkship.[2] Using a 9-item instrument with a 4-point scale, they found that seven ratings provided a reliable measure of an individual student's overall clinical skills. More observations are required for a reliable assessment of interpersonal aspects of clinical performance.

Although performance ratings are also used in residency training, faculty are the sole raters in many settings. There are exceptions. For many years at the University of Washington Department of Medicine, second-year and third-year residents have evaluated interns using the ABIM rating form. These ratings complement evaluations obtained from faculty, fellows, and chief residents. Mean ratings from all sources for each year are available to the program director for use in preparing a summary statement about the resident to submit to the ABIM. In addition, summative and normative data are presented to individual residents in two annual meetings—one with the program director or associate program director and one with the resident's continuity clinic mentor. Summative data without comments are also reviewed at quarterly meetings of the Clinical Competence Committee. Each resident may consult his or her file and use the ratings for self-evaluation and improvement. The files are confidential, and they may not be accessed without the resident's permission, except by staff in the residency office and for the purposes described above.

The study reported in this issue by Thomas and colleagues suggests that more extensive use of peer ratings may be on the horizon for residents.[3] This new study follows on the heels of another study examining the use of peer ratings in residency[4] and a review of the use of ratings scales in residency.[5] Thomas et al.'s study focuses on the assessment of interns by peer interns and residents, and compares these assessments with faculty assessments. In previous studies that have compared ratings of residents by attending physicians with those by other types of raters, faculty have often been more lenient raters. Explanations of leniency frequently pointed to lesser exposure of faculty to residents. Increased supervision requirements emanating from recent Medicare guidelines may be resulting in more contact between residents and faculty and therefore more discriminating ratings.

The apparent resistance among residents to participate in the peer rating process that Thomas and colleagues found is similar to findings in the study of Van Rosendaal and Jennett.[4] Especially among interns who are new to residency, the "colleague/friend factor" may influence the trainee's perception of the nature of peer ratings. Thomas et al.'s data suggest that senior residents may be less resistant to using peer ratings. This difference could be the result of increased self-confidence with the passage of time in training, becoming accustomed to the process, or discovering the usefulness of ratings for self-assessment. As those and other investigators have pointed out, training residents in evaluation and feedback should enhance the accuracy of and reactions to the ratings. In addition, as the use of peer ratings becomes more common in the practice setting, it is likely that residents will become more accepting of them as a routine part of training. More widespread use of ratings by peer residents as well as faculty should provide a link between use of performance ratings in medical school and peer ratings in the practice setting. This would result in a continuous assessment methodology throughout internists' medical careers.

Before residency-training programs begin widespread use of peer ratings, it is important to ensure that sound

methods are used to implement them. Thomas et al.'s pilot study provides grounds upon which to emphasize this point. According to the authors, their study confirms that peer review is reliable, feasible, and acceptable to residents, and that it provides different information than faculty assessments. However, the authors do not answer all the important questions about reliability. Although the Chronbach's α indicates high scale reliability, we do not know whether this measure of reliability is based on all items or on a subset of items. In Table 3 of their article, correlations are presented for only 8 of the 10 items. We wonder whether these 8 items were the basis for the tests of reliability in place of all 10 items. In addition, no information is presented about the reliability of the individual scale items.

Another important question that Thomas et al.'s study leaves unanswered is how many ratings are required to obtain a reliable rating of an individual resident. The authors provide no information about generalizability, which permits the estimation of the reliability of a measure using various numbers of observations. Generalizability should be addressed both for an overall rating and for individual scale items, as the two may have different results. Whether peer ratings are used in critical decisions, such as the annual assessment by the program director submitted to the ABIM, or for self reflection, the best measure is the most accurate one. Among practicing physicians rated by peer physicians with a form developed at the University of Washington in conjunction with the ABIM, 10 to 12 ratings provide a reliable assessment of overall clinical skills.[6,7] Among third-year medical students using a 9-item instrument with a 4-point scale, seven ratings provided a reliable rating of an individual student's overall clinical skills.[2] Similar studies should be done to estimate the number of ratings required to reliably assess residents' clinical skills. The domain of performance that is being examined will influence how many ratings are needed. Typically, assessments of interpersonal skills and professionalism require more ratings than assessment of cognitive skills. Another factor to consider is whether questions have been modified in any way. In Thomas et al.'s study, questions were modified somewhat from the ABIM rating form. When changes are made, it is preferable to reassess the psychometric characteristics of the instrument. Small changes in wording can have dramatic effects on meaning or result in ambiguity.

By making these caveats for moving ahead carefully, we are not arguing for delaying the use of peer ratings. Peer ratings were not taken seriously for many years because physicians assumed they were not reliable and did not provide a rigorous assessment of skills. And, in fact, when only one or two ratings were collected on an individual physician, this assumption was often found to be true. But since the introduction of careful psychometric methodology to the use of peer ratings, the reliability of this method of assessment has been established. With the recent ability to introduce increased efficiency in their implementation through electronic technologies, peer ratings represent a major advance in physician assessment. It is important to move ahead with the use of peer ratings with the care they require and deserve.—**Paul G. Ramsey, MD,** *and* **Marjorie D. Wenrich, MPH,** *University of Washington School of Medicine, Seattle, Wash.*

REFERENCES

1. ABIM Board of Directors. Report of the ABIM Task Force on Recertification: A Program for Continuous Professional Development. Philadelphia, Pa: ABIM; June 1999.
2. Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of student's clinical skills in a medicine clerkship. J Gen Intern Med. 1992;7:506–10.
3. Thomas PA, Gebo KA, Hellmann DB. A pilot study of peer review in residency training. J Gen Intern Med. 1999;14:551–4.
4. Van Rosendaal GM, Jennett PA. Comparing peer and faculty evaluations in an internal medicine residency. Acad Med. 1994;69:299–303.
5. Gray JD. Global rating scales in residency education. Acad Med. 1996;71:S55–62.
6. 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.
7. Ramsey PG, Carline JD, Blank LL, Wenrich MD. Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. Acad Med. 1996;71:364–70.