JAMA
The Journal of the American Medical Association

Volume 269(13), 7 Apr 1993, pp 1655-1660

Use of Peer Ratings to Evaluate Physician Performance

[Original Contributions]

Ramsey, Paul G.; Wenrich, Marjorie D.; Carline, Jan D.; Inui, Thomas S.; Larson, Eric B.; LoGerfo, James P.

From the Departments of Medicine (Drs Ramsey, Larson, Inui, and LoGerfo, and Ms Wenrich) and Medical Education (Dr Carline), University of Washington School of Medicine, Seattle.
Reprint requests to Department of Medicine, RG-20, University of Washington School of Medicine, Seattle, WA 98195 (Dr Ramsey).

Abstract

OBJECTIVE: To assess the feasibility and measurement characteristics of ratings completed by professional associates to evaluate the performance of practicing physicians.

DESIGN: The clinical performance of physicians was evaluated using written questionnaires mailed to professional associates (physicians and nurses). Physician-associates were randomly selected from lists provided by both the subjects and medical supervisors, and detailed information was collected concerning the professional and social relationships between the associate and the subject. Responses were analyzed to determine factors that affect ratings and measurement characteristics of peer ratings.

SETTING AND PARTICIPANTS-Physician-subjects were selected from among practicing internists in New York, New Jersey, and Pennsylvania who received American Board of Internal Medicine certification 5 to 15 years previously.

MAIN OUTCOME MEASURE: Physician performance as assessed by peers.

RESULTS: Peer ratings are not biased substantially by the method of selection of the peers or the relationship between the rater and the subject. Factor analyses suggest a two-dimensional conceptualization of clinical skills: one factor represents cognitive and clinical management skills and the other factor represents humanistic qualities and management of psychosocial aspects of illness. Ratings from 11 peer physicians are needed to provide a reliable assessment in these two areas.

CONCLUSIONS: These findings suggest that it is feasible to obtain assessments from professional associates of practicing physicians in areas such as clinical skills, humanistic qualities, and communication skills. Using a shorter version of the questionnaire used in this study, peer ratings provide a practical method to assess clinical performance in areas such as humanistic qualities and communication skills that are difficult to assess with other measures.

(JAMA. 1993;269:1655-1660)

Evaluation of the performance of practicing physicians has received increased attention recently from many diverse groups. This attention stems from several sources, including the speed with which new medical information is developed and disseminated, data suggesting that medical knowledge declines over time, and public interest in the credentialing and licensure processes as potential methods of ensuring excellent medical care [1,2,3]. Interest in quality of clinical performance has also broadened in recent years to include evaluation of physicians' communication skills and humanistic qualities [4,5,6]. Despite increased attention to physician assessment, most methods used for this purpose either measure limited or unidimensional aspects of clinical performance or are too complex to be feasible to assess performance of a large number of individual physicians [3,7].

One method of physician evaluation, global performance ratings, has been used frequently in the medical education setting to evaluate the performance of medical students and residents. Faculty members rate students and residents in clinical skills, humanistic qualities, and communication skills based on repeated observations over a period of time. The reliability of global performance ratings in the training setting has been explored, and global ratings can provide a reliable assessment of medical students' performance if a sufficient number of observers is used [8].

Performance ratings have also been used to evaluate practicing physicians as a part of the credentialing processes for medical societies and hospital privileges. Hospital and county medical societies commonly obtain ratings concerning candidates from a few professional associates. Recently, peer ratings have been considered as an evaluation mechanism for recertification of practicing physicians by specialty boards, including the American Board of Internal Medicine (ABIM) [1]. However, little information is available about the measurement characteristics of peer ratings in the practice setting. A recent study of the predictive validity of certification by the ABIM was the first large-scale study to use global ratings to assess practicing physicians' performance [9,10], but results from this study were not sufficient to explore the factors that might affect these ratings.

The study described here was designed to assess the feasibility and measurement characteristics of ratings completed by professional associates to assess humanistic qualities, communication skills, and selected aspects of clinical skills of practicing internists. Two methods of selecting physician-associates (selection by the subject and by a medical supervisor) were used to permit analysis of possible bias associated with selection source, and extensive demographic information was obtained concerning the physician-associates and the degree of professional and social involvement with the physician-subject. The results suggest that peer ratings can provide an assessment of practicing internists' overall clinical skills, communication skills, and humanistic qualities that might be used in conjunction with other performance measures.

METHODS

Study Sample and Selection Criteria

Potential physician-subjects in New York, New Jersey, and Pennsylvania were identified from data provided by the American Board of Medical Specialties and were approached by a recruitment letter. Physicians were chosen as subjects from among internists who had been certified in internal medicine 5 to 15 years previously in order to explore the feasibility of peer ratings to assess the performance of internists as part of an ABIM recertification process [11]. Approximately two thirds of the group were selected from general internists who served as the principal physician for at least 60% of their patients. The remaining subjects were selected from among internal medicine subspecialists. Potential physician-subjects were excluded from participating in the study if the physician had moved out of state, spent less than 60% of his or her professional time in clinical activity, or spent more than 40% of his or her professional time practicing another specialty (eg, family medicine). The study was reviewed and approved by the University of Washington Human Subjects Review Committee.

Evaluation Instruments and Strategies

An extensive demographic questionnaire for completion by physician-subjects was designed to obtain information about subjects' training background, certification, practice characteristics, faculty appointments, and professional and educational activities. The names of admitting hospitals, medical supervisors, and physicians from selected specialties with whom the physician-subjects worked in patient care in the last year were also obtained. In addition, physicians were asked to complete a brief survey concerning their opinions about the potential uses of peer ratings.

Questionnaires used by physician-associates in a previous study of the predictive validity of certification by the ABIM [9] were revised and expanded to include questions related to humanistic qualities, communication skills, use of laboratory tests and diagnostic procedures, and inpatient and outpatient clinical management skills. A nine-point Likert scale was used to obtain ratings in categories including verbal communications, management prior to referral, medical knowledge, integrity, psychosocial aspects of illness, management of multiple complex problems, responsibility, and overall clinical skills. The questionnaire was also used to elicit demographic data about the professional associate and about prior professional and social contacts with the physician-subject. A questionnaire was also developed for completion by nurse-associates identified by head nurses on floors of hospitals to which subjects admitted patients.

Identification and Recruitment of Physician-Associates

Physician-associates were defined as physicians to whom the physician-subject referred patients, with whom responsibility for patients had been shared, or from whom consultations had been obtained in the last year. Two different methods for selecting physician-associates were used. One group of associates was selected by the physician-subject. Using a list of associates generated by the subject, a stratification procedure was employed to favor the selection of internists, neurologists, and dermatologists. The second method used to identify physician-associates involved identification of a physician in a supervisory position (eg, chief of medicine, chief of staff, or medical director) for each physician-subject. The supervisor was asked to provide a list of physician-associates to complete ratings on the physician-subject. The same stratification procedure was used to select physician-associates from this list provided by the medical supervisor as for the list of physician-associates provided by the subject. A maximum of 25 associates were randomly selected from the lists provided by the physician-subject and the supervisor. Associates selected from both lists were identified by a unique code. These associates completed only one questionnaire, but their responses were in- cluded in both sets of data.

Physician-Associates Identified by Subjects

The 316 subjects (98.7%) who pro- vided lists of associates named a mean of 33.1 associates. A total of 314 subjects agreed finally to participate and provided at least one identifiable associate name. Average numbers of associates named by subjects in different community sizes were similar (metropolitan, 34; cities, 33; towns, 33). Addresses and demographic information were available for a mean of 20.7 associates per subject. A questionnaire and cover letter signed by the subject were mailed to each selected physician. As many as two follow-up requests were mailed to nonrespondents. A mean of 8.7 questionnaires was returned per subject from physician-associates who appeared to be qualified to provide ratings. After exclusion of ineligible associates (those for whom mailing addresses appeared to be incorrect or who returned a questionnaire indicating that they did not work with the physician in question), the response rate from physician-associates named by the subjects was 51.6%. The response rate was similar for subjects in metropolitan areas (51.8%), cities (51.5%), and towns (51.7%).

Physician-Associates Identified by Medical Supervisors

Medical supervisors named by subjects were asked to independently generate a list of physicians who may have worked with the subject in patient care or to send hospital rosters from which names could be selected. Of the medical supervisors identified by 309 physician-subjects, 205 returned lists of associates or rosters, with a mean of 36.7 associates named per physician-subject. Addresses were found and demographic information was available for a mean of 17.3 associates per subject. A questionnaire and cover letter signed by the subject were mailed to each associate, and as many as two follow-ups were mailed to nonrespondents. A mean of 7.3 questionnaires was returned per subject from lists generated by medical supervisors. After exclusion of ineligible physicians and those with inaccurate addresses, the overall response rate for associates from supervisor-generated lists was 52.1%.

Comparison of Participating Physician-Subjects With Nonparticipants

A sample of 500 randomly selected physicians who had been recruited but who did not participate in the project was selected to compare with participants to determine the representativeness of the study sample. Prior ABIM certifying examination scores and demographic data obtained from questionnaires and medical directories were compared for participants and nonparticipants. No significant differences in ABIM certifying examination scores were found between the participants and nonparticipants. The mean score for participants on the ABIM certifying examination (509.4 77) was similar to the mean score for nonparticipants (502.0 76.6) (P=not significant). Training background, type of practice, and other demographic characteristics were also compared for participants and randomly selected nonparticipants. Using medical school rankings from a published list [12], there were no differences in medical school training programs found between participants and nonparticipants. Furthermore, the nonparticipants were as likely to have trained in a university-based residency program as the participating physician-subjects. Thus, based on these data, participants appeared to be representative of practicing certified internists in New York, New Jersey, and Pennsylvania who received certification 5 to 15 years previously.

Statistical Methods

Initial investigation of the associate ratings centered on generalizability studies using analysis of variance (ANOVA) methods described by Winer [13]. Rating items that received many "unable to evaluate" responses or missing responses were excluded from further analysis. A minimum generalizability coefficient of .7 was selected to identify physicians whose performance was considerably above or below their peers. Subject-level ratings were calculated that represent the mean rating across all raters for that subject, and direct analyses of the ratings were completed at the subject level. Direct analyses of the effects on ratings of different methods for selecting physician-associates were made at the subject level. Subject-level scores were calculated separately for associates named by the subject and those named by the medical supervisor. Associates selected from both sources were included in both ratings. Using subject-level ratings, factor analyses of the ratings using Varimax factor rotation were completed to identify internal relationships in the ratings and to establish subscales of ratings. The relationships between nurse-associate and physician-associate ratings for those items that were common to the two questionnaires were investigated using Pearson product moment correlations. The relationships of subject demographic characteristics to ratings were determined using Pearson product moment correlations, Student's t test, chi squared (chi^2), and one-way ANOVA, where appropriate. Comparisons of associate ratings by demographic characteristics of the raters at the rater level were completed using these same statistical tests. Exploratory stepwise multiple regression was performed to determine associate-based predictors of ratings of overall clinical skills by physician-associates. The alpha level for a significant difference was set at P<.05.

RESULTS

Training and Practice Characteristics of Physician-Subjects

The final sample of ABIM-certified internists consists of 147 physician-subjects who live or practice in New York, 115 subjects in Pennsylvania, and 56 subjects in New Jersey. Subjects were equally divided among communities of different sizes (106 each in metropolitan areas, cities, and towns). Among the 318 subjects, 41 (12.9%) were women. A total of 73 subjects (23%) graduated from foreign medical schools. Mean year for completion of residency training was 1978 (SD, 4.4), and mean year of ABIM certification was 1979 (SD, 3.3). Approximately 60% of the subjects had fellowship training following residency, and 151 (47.5%) of subjects were board-certified in a subspecialty. Approximately two thirds of the subjects indicated that they served as principal physician for at least 60% of their patients.

Characteristics of Physician-Associates

Most associates (91.8%) who returned questionnaires were men. The mean age among responding associates was 45.4 years (SD, 8.7). One third of the respondents were in solo practice, and 47% were in a private practice with at least one other physician. Approximately 79% of the physician-associates practiced general internal medicine, a subspecialty of internal medicine, dermatology, or neurology. The most common subspecialty practices among the physician-associates were cardiology, gastroenterology, hematology/oncology, pulmonary and critical care medicine, and nephrology.

Some differences in the nature and amount of contact between physician-subjects and associates were found based on whether the associates were selected by the subject or by a medical supervisor. The associates selected by the medical supervisor had more extensive contacts with the subjects in professional non-patient-related contacts, social contacts, and interactions observed with nurses and hospital personnel. By contrast, physician-associates selected by the subject reported more patient referrals from the subject. The associates selected by the subject and by the supervisor indicated the same number of patients referred from the associate to the subject. There were no differences between number of contacts with associates from the two list sources in number of inpatients and outpatients shared, inpatient and outpatient interactions observed, interactions with patients' families observed, and professional patient-related contacts in general.

Effects of the Method of Selecting Physician-Associates on Ratings

Ratings by physician-associates selected by the subject were compared with ratings by physician-associates selected by a medical supervisor using subject-level data in a paired t test analysis. Among the 20 categories on the physician-associate questionnaire, there were only two small differences found between the ratings of the associates selected by the two different methods, and these differences would not be significant if Bonferonni's correction for multiple tests was applied.

A second approach to determining the effects of the method of selecting raters was to use multiple regression techniques to estimate the variance of ratings accounted for by rater sources. Although the variance contribution of the source of the ratings in three questionnaire categories was statistically significant (respect, psychosocial aspects of illness, and compassion), the actual variances were small. Variance estimates for all categories ranged from .0001 to .0045. From these analyses and the direct comparisons of ratings for subjects, it appears that the method of selecting the physician-associates had a negligible effect on ratings. Therefore, ratings from associates from all sources were used for subsequent analyses.

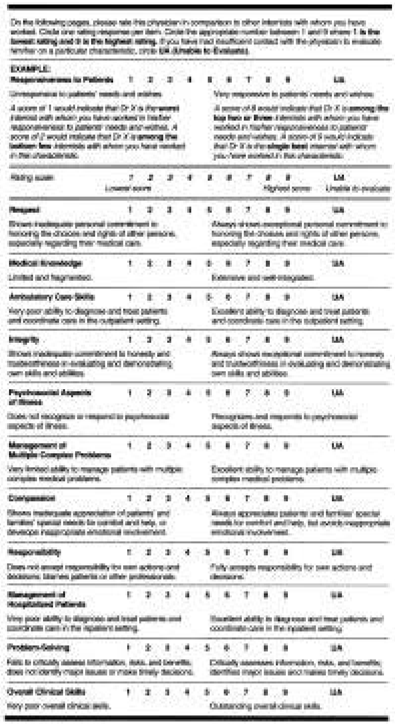
Physician-Associate Ratings

Physician-associate ratings were available for 313 of the certified physician-subjects. The mean ratings for the categories evaluated ranged from a low of 7.4 for verbal communications to a high of 8.1 for integrity. Standard deviations ranged from a low of 1.1 for the rating of integrity to a high of 1.6 for the rating of whether the associate would refer a family member to the physician-subject for care. More than 90% of associates provided a rating for the majority of categories. Intercorrelations between items on the physician-associate questionnaire ranged from .47 (ratings of medical knowledge and compassion) to .86 (ratings of problem-solving skills and management of hospitalized patients, and ratings of problem-solving and overall clinical skills).

Preliminary analyses indicated that 13 ratings would be required to obtain generalizability coefficients of .7 or greater for individual items. To assure an overall mean number of 13 responses per subject for further analyses, subjects with fewer than seven associate ratings were excluded from further generalizability calculations. Generalizability coefficients of at least .7 can be obtained with 13 responses for a majority (69%) of individual questionnaire items. Mean questionnaire ratings were calculated for subjects who had obtained a sufficient number of ratings to achieve this level of reliability, and these subject-level ratings were used for later analyses. Ratings of laboratory skills were associated with low generalizability coefficients and many "unable to evaluate" responses and were excluded from further analyses.

To determine if patterns of relationships between questionnaire items existed, a principal components factor analysis was conducted on subject-level ratings. With

Varimax rotation, two factors were identified that accounted for 86% of the variance in the ratings. The first factor received high factor loadings on items that represented cognitive and clinical management skills. The second factor received high factor loadings on items representing humanistic qualities, responsibility, and management of psychosocial aspects of illness. Using the results of the factor analysis and an investigation of the intercorrelations, a second analysis of the data using a reduced number of items was undertaken. Five items from each of the two factors along with the overall clinical skills item were resubmitted to principal components factor analysis. This reduced set of items resulted in two factors accounting for 89% of the variability Table 1. Generalizability studies were performed to estimate the number of physician-associate ratings needed to achieve reliable ratings for each of these factors in the reduced data set. To achieve a generalizability (reliability) coefficient of at least .7, 11 physician-associate ratings would be required for each area. For the rating of overall clinical skills, 10 physician-associate ratings were required to achieve a generalizability coefficient of .7.



On the following pages, please rate this physician in comparison to other internists with whom you have worked. Circle one rating response per item. Circle the appropriate number between 1 and 9 where **1 is the lowest rating and 9 is the highest rating.** If you have had insufficient contact with the physician to evaluate him/her on a particular characteristic, circle **UA (Unable to Evaluate).**

EXAMPLE:

Responsiveness to Patients	1	2	3	4	5	6	7	8	9	UA
Unresponsive to patients' needs and wishes.					Very responsive to patients' needs and wishes.					
A score of 1 would indicate that Dr X is the **worst** internist with whom you have worked in his/her responsiveness to patients' needs and wishes. A score of 2 would indicate that Dr X is **among the bottom few** internists with whom you have worked in this characteristic.					A score of 8 would indicate that Dr X is **among the top two or three** internists with whom you have worked in his/her responsiveness to patients' needs and wishes. A score of 9 would indicate that Dr X is the **single best** internist with whom you have worked in this characteristic.					
Rating scale:	1 Lowest score	2	3	4	5	6	7	8	9 Highest score	UA Unable to evaluate
Respect	1	2	3	4	5	6	7	8	9	UA
Shows inadequate personal commitment to honoring the choices and rights of other persons, especially regarding their medical care.					Always shows exceptional personal commitment to honoring the choices and rights of other persons, especially regarding their medical care.					
Medical Knowledge	1	2	3	4	5	6	7	8	9	UA
Limited and fragmented.					Extensive and well-integrated.					
Ambulatory Care Skills	1	2	3	4	5	6	7	8	9	UA
Very poor ability to diagnose and treat patients and coordinate care in the outpatient setting.					Excellent ability to diagnose and treat patients and coordinate care in the outpatient setting.					
Integrity	1	2	3	4	5	6	7	8	9	UA
Shows inadequate commitment to honesty and trustworthiness in evaluating and demonstrating own skills and abilities.					Always shows exceptional commitment to honesty and trustworthiness in evaluating and demonstrating own skills and abilities.					
Psychosocial Aspects of Illness	1	2	3	4	5	6	7	8	9	UA
Does not recognize or respond to psychosocial aspects of illness.					Recognizes and responds to psychosocial aspects of illness.					
Management of Multiple Complex Problems	1	2	3	4	5	6	7	8	9	UA
Very limited ability to manage patients with multiple complex medical problems.					Excellent ability to manage patients with multiple complex medical problems.					
Compassion	1	2	3	4	5	6	7	8	9	UA
Shows inadequate appreciation of patients' and families' special needs for comfort and help, or develops inappropriate emotional involvement.					Always appreciates patients' and families' special needs for comfort and help, but avoids inappropriate emotional involvement.					
Responsibility	1	2	3	4	5	6	7	8	9	UA
Does not accept responsibility for own actions and decisions; blames patients or other professionals.					Fully accepts responsibility for own actions and decisions.					
Management of Hospitalized Patients	1	2	3	4	5	6	7	8	9	UA
Very poor ability to diagnose and treat patients and coordinate care in the inpatient setting.					Excellent ability to diagnose and treat patients and coordinate care in the inpatient setting.					
Problem-Solving	1	2	3	4	5	6	7	8	9	UA
Fails to critically assess information, risks, and benefits; does not identify major issues or make timely decisions.					Critically assesses information, risks, and benefits; identifies major issues and makes timely decisions.					
Overall Clinical Skills	1	2	3	4	5	6	7	8	9	UA
Very poor overall clinical skills.					Outstanding overall clinical skills.					

Table 1. No caption available

[Help with image viewing]
[Email Jumpstart To Image]

Effects of the Relationship Between Physician-Associates and Subjects on Ratings

The effects of the relationship between physician-associates and physician-subjects on ratings were investigated using several indicators of the nature and amount of contact over the last 6 months, including number of outpatients shared, number of inpatients shared, the total number of professional patient-related contacts, the number of patients referred from the associate to the subject, and the number of patients who were referred by the subject to the associate. Significant positive correlations were found between the numbers of patients shared or referred and the peer ratings. However, all correlation coefficients were low (correlation coefficients range from .08 to .30). The highest correlations found were between the number of patients referred by the associate to the subject and the peer ratings. All other correlations were generally less than .20. Stepwise regression analyses were also performed to determine the extent to which the relationship between physician-associates and subjects influenced overall ratings. Variables related to the associate's relationship with the subject were identified by correlational analyses. Although several predictor variables were statistically significant, the total amount of variability explained was only 9%. The number of referrals by the associate to the subject accounted for more than 75% of the explained variability.

Correlation Between Physician-Associate and Nurse-Associate Ratings

Observed and disattenuated correlations were calculated for the 10 categories common to both the physician-associate and nurse-associate rating forms. Disattenuated correlation coefficients that are predictions of the correlation between instruments if the instruments were perfectly reliable, ranged from a low of .46 for responsibility to a high of .72 for compassion. Disattenuated correlation coefficients greater than .5 were seen for all categories except responsibility.

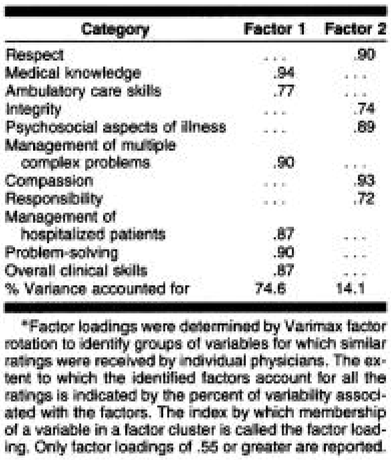
Physicians' Opinions About Peer Ratings

Of 263 physician-subjects who completed a survey concerning the possible use of peer ratings, 84% felt that ratings completed by physicians should be used to evaluate the overall clinical skills of practicing internists for such purposes as credentialing and recertification. Only 7.3% of the physician-subjects felt that peer ratings should not be used, and the remaining 8.4% expressed uncertainty about use of peer ratings for these purposes. Approximately three quarters of the physician-subjects also felt that peer ratings by physicians should be used to evaluate the humanistic qualities and communication skills of practicing physicians.

COMMENT

Increased public interest in evaluation of the performance of practicing physicians has created a need for evaluation strategies that provide a reliable and comprehensive assessment of individual physician performance. Written examinations are the most widely used evaluation method, but written examinations do not provide information about areas of performance such as communication skills and humanistic qualities that have important effects on patient care. Peer ratings offer the advantage of being performance-based, and these ratings may provide a useful evaluation of performance in areas that are difficult to assess reliably with other methods. However, the measurement characteristics of peer ratings in the practice setting have not been studied. Critics of this technique argue that negative ratings by one or two physicians can unfairly influence a physician's overall evaluation and that peer ratings are unreliable. Other potential limitations of peer ratings in the practice setting include bias associated with selection of raters by the physician being evaluated, the effects on ratings of the working relationship of the evaluator to the physician being evaluated, and the presence of other factors that may affect ratings.

The current study was designed to explore the measurement characteristics of peer ratings used to assess the clinical skills, humanistic qualities, and communication skills of practicing physicians. The reliability and feasibility of peer ratings were assessed, and factor analyses were performed to determine relationships among the areas of performance examined. Data obtained from principal component factor analysis in this project support a two-dimensional conceptualization of clinical skills. The first factor includes items related to medical knowledge, problem-solving skills, management of complex patients, and overall clinical skills. The second factor identified includes items concerning humanistic qualities (respect, integrity, and compassion) and items related to responsibility and management of the psychosocial aspects of illness. When a shortened form is used Figure 1, results suggest that reliable rat ings for identifying physicians whose performance is considerably above or below their peers can be achieved for each factor with 11 responses. Because these data also suggest that the method of selecting associates and the relationship between the rater and the subject do not substantially bias results, evaluators can be identified by the physician to be rated.



Category	Factor 1	Factor 2
Respect	. . .	.90
Medical knowledge	.94	. . .
Ambulatory care skills	.77	. . .
Integrity	. . .	.74
Psychosocial aspects of illness	. . .	.89
Management of multiple complex problems	.90	. . .
Compassion	. . .	.93
Responsibility	. . .	.72
Management of hospitalized patients	.87	. . .
Problem-solving	.90	. . .
Overall clinical skills	.87	. . .
% Variance accounted for	74.6	14.1

*Factor loadings were determined by Varimax factor rotation to identify groups of variables for which similar ratings were received by individual physicians. The extent to which the identified factors account for all the ratings is indicated by the percent of variability associated with the factors. The index by which membership of a variable in a factor cluster is called the factor loading. Only factor loadings of .55 or greater are reported.

Figure 1.

[Help with image viewing]
[Email Jumpstart To Image]

Peer ratings appear to be a feasible method. Physician-subjects in all geographic locations (metropolitan, city, town) identified an adequate mean number of associates to achieve the required number of ratings for reliable results with a 50% response rate. A previous study undertaken on the West Coast also suggests that sufficient numbers of associates could be identified to achieve adequate numbers of ratings for reliable results [10]. The response rates obtained in the current study, however, can probably be considered at the low end of possible response rates. The length of the questionnaire used (five pages) and the inclusion of personal questions probably reduced the likelihood of responses. In a previous study, a one-page rating form was used with no follow-up requests and the response rate from internists and surgeons was greater than 75% [9]. Since factor analyses in this study indicate that the number of questions can be reduced to 11 items and still maintain the factorial structure of the original data, a shorter form can be used that would probably result in higher response rates. Peer ratings also appear to be an evaluation technique that is acceptable to practicing physicians. Eighty-four percent of subjects in this study support the use of peer ratings of clinical skills for purposes such as credentialing and recertification.

Although the absence of criterion standards makes it difficult to study the validity of assessment measures for individual physicians, comparison of the results of the physician-associate ratings with results from other measures of excellence suggests that peer ratings are meaningful. Strong correlations in the range of .5 to .6 have been found between peer ratings of medical knowledge and ABIM examination scores, and low correlations (<.15) were found between ratings of humanistic qualities and examination scores [9]. Although there are no established criterion measures for assessing humanistic qualities and communication skills, use of ratings in these areas by registered nurses provides an independent measure to compare with the physician-associate ratings. There was relatively strong agreement in the ratings of humanistic qualities, communication skills, and selected aspects of clinical skills between physician-associates and nurse-associates. Further work is needed to determine whether peer ratings predict patient outcomes, and additional work is also needed to explore the use of peer ratings in specialties other than internal medicine.

Peer ratings may be the oldest and most commonly used informal method for physicians to assess the performance of a colleague. For example, when a physician seeks to identify another physician in a different geographic area to care for a patient, friend, or family member, he or she often obtains one or more peer ratings from other physicians in that community. Increasingly, the types of questions used in peer ratings are used formally as well as informally. The global question, "Would you send your mother to these physicians in this hospital?" is used now as part of quality of care implicit reviews [14]. Despite the apparent "face validity" of peer ratings, this evaluation strategy has received little systematic attention due to the perception that these ratings are unreliable and represent little more than personal recommendations from friends. Based on the typical practice used for collecting peer ratings, this perception is true. The results from our project suggest that the small numbers of professional associates that are typically used in peer assessments obtained by many hospitals or county medical society credentialing processes are insufficient to obtain reliable information. However, our results show that reliable ratings can be obtained with 11 responses and that it is feasible to obtain the number of ratings needed to provide a reliable assessment of an individual physician's performance for use in identifying outlying physicians. Although further work is needed to confirm these results outside the research setting, our findings suggest that peer ratings provide a practical method to assess the performance of practicing internists in areas such as humanistic qualities and communication skills that are difficult to evaluate reliably with other measures.

This study was supported by the American Board of Internal Medicine and by a Henry J. Kaiser Family Foundation Faculty Scholar Award in General Internal Medicine (Dr Ramsey). The article does not necessarily reflect the opinions or policies of the American Board of Internal Medicine or the Henry J. Kaiser Family Foundation.

The investigators greatly appreciate the support and advice of John Benson, MD, George Webster, MD, Susan Day, MD, John Norcini, PhD, David Swanson, PhD, and Lou Grosso, MS. Donald Langsley, MD, at the American Board of Medical Specialties provided valuable data for recruitment of subjects. Lois Giles provided excellent support in all aspects of the project, and Martha Mihara and Jill Orman provided support in data collection.

REFERENCES

1. Glassock RJ, Benson JA, Copeland RB, et al. Time-limited certification and recertification: the program of the American Board of Internal Medicine. Ann Intern Med. 1991;114:59-62. **Bibliographic Links** [Context Link]

2. Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. Changes over time in the knowledge base of practicing internists. JAMA. 1991;266:1103-1107. **Bibliographic Links** [Context Link]

3. Langsley DG. Medical competence and performance assessment. JAMA. 1991;266:977-980. **Bibliographic Links** [Context Link]

4. Keynan A, Friedman M, Benbassat J. Reliability of global rating scales in the assessment of clinical competence of medical students. Med Educ. 1987;21:477-481. **Bibliographic Links** [Context Link]

5. Blurton RR, Mazzaferri EL. Assessment of interpersonal skills and humanistic qualities in medical residents. J Med Educ. 1985;60:648-650. **Bibliographic Links** [Context Link]

6. Linn LS, DiMatteo MR, Cope DW, Robbins A. Measuring physicians' humanistic attitudes, values, and behaviors. Med Care. 1987;25:504-513. **Bibliographic Links** [Context Link]

7. Neufeld VR, Norman GR, eds. Assessing Clinical Competence. New York, NY: Springer Publishing Co; 1985. [Context Link]

8. Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. J Gen Intern Med. 1992;7:506-510. **Bibliographic Links** [Context Link]

9. Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JB, Wenrich MD. Predictive validity of board certification by the American Board of Internal Medicine. Ann Intern Med. 1989;110:719-726. **Bibliographic Links** [Context Link]

10. Carline JD, Wenrich MD, Ramsey PG. Characteristics of ratings of physician competence by professional associates. Eval Health Prof. 1989;12:409-423. [Context Link]

11. Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich M. Final Report to the American Board of Internal Medicine: Assessment of the Clinical Competence of Certified Internists. Philadelphia, Pa: American Board of Internal Medicine; 1990. [Context Link]

12. Gourman J. The Gourman Report: A Rating of Graduate and Professional Programs in American and International Universities. Los Angeles, Calif: National Education Standards Inc; 1980:64-66. [Context Link]

13. Winer B. Statistical Principles in Experimental Design. 2nd ed. New York, NY: McGraw Hill International Book Co; 1971:283-296. [Context Link]

14. Keeler EB, Rubenstein LV, Kahn KL, et al. Hospital characteristics and quality of care. JAMA. 1992;268:1709-1714. **Bibliographic Links** [Context Link]