A Pilot Study of Peer Review in Residency Training

Patricia A. Thomas, MD, Kelly A. Gebo, MD, David B. Hellmann, MD

OBJECTIVE: **To explore the utility of peer review (review by fellow interns or residents in the firm) as an additional method of evaluation in a university categorical internal medicine residency program.**

DESIGN/PARTICIPANTS: **Senior residents and interns were asked to complete evaluations of interns at the end-of-month ward rotations.**

MAIN RESULTS: **Response rates for senior residents evaluating 16 interns were 70%; for interns evaluating interns, 35%. Analysis of 177 instruments for 16 interns showed high internal consistency in the evaluations. Factor analysis supported a two-dimensional view of clinical competence. Correlations between faculty, senior resident, and intern assessments of interns were good, although varied by domain.**

CONCLUSIONS: **An end-of-year attitude survey found that residents gave high ratings to the value of feedback from peers.**

KEY WORDS: **education; internal medicine; peer evaluation; residency training; humanism.**
J GEN INTERN MED 1999;14:551–554.

Modern training programs struggle to find reliable methods of evaluating the humanistic aspect of clinical competence.[1–3] Busy faculty may have limited opportunities to observe residents interacting with patients, and patient surveys are costly, requiring as many as 30 to 40 surveys to provide reliable assessment.[4] Peer review in medical education has been shown to be reliable and to add unique information to the assessment of trainees.[5–8] In studies of practicing physicians, peer assessments were found to be reliable and generalizable with 6 to 11 responses, and were well accepted.[9–11]

As an evaluation method in residency training, peer review should complement faculty and objective assessments already in use. Peers in ward teams have unique opportunities to observe the professional behaviors of their colleagues. In addition, the process of peer review promotes personal skills of self-assessment and feedback. Inclusion into the peer instruments of those domains that are valued by the program, such as integrity, teamwork, and teaching skills, focuses resident attention to these domains. Peer assessment has been incorporated into the American Board of Internal Medicine (ABIM) recertification process, and experience with this form of assessment should become part of the training for future professional life.[12,13]

Despite these advantages, peer assessment is rarely included in resident evaluation, suggesting significant barriers to its use.[14,15] Residents work in stressful environments and rely on mutual support to cope with stress during the training process. They may resist the use of peer review for this reason, or rate their colleagues on the basis of friendships rather than specific observations, resulting in evaluations with low validity. It is also unclear whether residents would value the anonymous opinions of colleagues to the point of altering behavior.

We hypothesized that peer assessment of interns would provide information different from that provided by faculty assessments, especially in the areas of humanistic and professional behaviors, and we sought to explore the issues of feasibility, reliability, and resident reaction to the use of peer review through a pilot intervention.

METHODS

Two of four inpatient firms were chosen to pilot test peer assessment. The inpatient firms rotate monthly in teams of four interns, two senior residents, a chief resident, and one teaching attending. The peer review instrument was constructed with 10 items to reflect the domains of the ABIM evaluation form,[16] with additions suggested by residents. A 9-point global rating scale was used for each item (Appendix A).

At the end-of-month ward rotations, interns were asked to complete evaluations of other interns and senior residents on the firm; senior residents completed evaluations of interns only. This report focuses on evaluations of interns by interns and senior residents. It was explained that forms would be anonymously collated before being returned to individual interns or residents, and that results would not be included in permanent resident files. Most senior residents had received training in feedback skills as part of a teaching skills course for rising seniors; interns were given no specific training in feedback and evaluation.

An attitude survey was mailed to all housestaff at the end of the pilot test year, in which residents rated the value of feedback to them from different types of evaluators, including teaching faculty, peers, medical students, other health professionals and patients.

Statistical analysis was performed with Simstat software. Kruskal-Wallis one-way analysis of variance was used to test differences between groups. Since the rating scales exhibited a ceiling effect, differences between groups were tested with bootstrap simulation. Factor analysis (principal component analysis) was used to determine which items in the instrument were related.

Received from the Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Md (PAT, KAG, DBH).

Address correspondence and reprint requests to Dr. Thomas: 1830 East Monument St., Room 9033, Baltimore, MD 21205.

RESULTS

During the 9-month pilot, 117 instruments were returned for 16 interns; 101 of these intern evaluations were completed by senior residents, and 76 were completed by interns. There were 72 intern months (4 interns per firm × 2 firms × 9 months). Thus, the response rate for interns evaluating interns was 35% (76/216); the response rate for senior residents evaluating interns was 70% (101/144). The two firms differed in their use of peer review; firm A returned an average of 13.8 evaluations per intern (range 8–23), and firm B returned an average of 7.1 evaluations per intern (range, 4–13), $p < .01$. For each intern, a sum score by item was calculated and averaged by the number of evaluations. Differences between firms were not statistically significant. A summary of intern evaluations by residents, interns, and faculty is show in Table 1.

Interrater reliability could not be determined because the returns were anonymous. For senior resident evaluations of interns, the average interitem correlation was .55, Cronbach's $\alpha = .93$. For intern assessments of interns, the average interitem correlation was .73, Cronbach's $\alpha = .96$.

Factor analysis (principal component analysis) was used to determine which items in the instrument were related. Two factors were identified. Factor 1, termed "technical skills," was weighted with those items representing cognitive and psychomotor skills and behaviors, and factor 2, "interpersonal skills," was weighted with items representing interpersonal skills and humanistic behaviors. Loadings in the Varimax rotation for senior residents and interns evaluating interns are listed in Table 2.

Sum peer evaluations for each intern were compared with similar items in the faculty and chief resident end-of-month evaluations (Table 3). These 16 interns had accumulated 197 faculty inpatient evaluations, an average of 12.3 evaluations per intern. There was good correlation between the two forms of evaluation. Senior resident and faculty correlations were above .60 in medical knowledge, history taking, procedural skills, clinical judgment, and overall competence. A different pattern was seen in correlations between senior resident and intern assessments of interns. The only correlation above .60 was for procedural skills. Correlations between faculty and intern evaluations were moderate to high except for medical knowledge.

All house officers in the program were asked to rate on a scale of 1 = none to 5 = extremely valuable, the value of feedback from peers in the traditional ABIM domains of clinical competence. Residents in the two firms exposed to peer review rated the value of this form of feedback slightly higher than residents in those firms not exposed to peer review, especially in the domains of medical knowledge, medical care, and moral and ethical behavior (range, 4.2–4.67 vs 3.56–4.33, Kruskal-Wallis $p < .03$).

DISCUSSION

Our study confirms that peer review is reliable, feasible (at least when done by residents), provides somewhat different information than faculty assessments, and is acceptable to residents.

Two concerns were raised in the pilot study that will challenge the value of peer review in a residency evaluation system: the response rate and the unknown criteria by which residents rated their peers. Our trainees, like practicing physicians who studied elsewhere, demonstrate a two-dimensional view of clinical competence when evaluating their peers: technical skills and interpersonal skills.[9] It is interesting that the variance in these two factors differed for the type of evaluator, but given the low response rate for interns, further study is needed to confirm this finding and understand its significance. Interns may be using different criteria or values in their assessments of their colleagues, or have different observational data. This was further suggested by the low correlations between senior

Table 1. Ratings of Interns by Residents and Interns*

Item	Mean Ratings of Interns by Residents (SD) (n = 101 evaluations)	Mean Ratings of Interns by Interns (SD) (n = 70 evaluations)	Mean Ratings by Faculty (n = 197 evaluations)
Medical knowledge	8.19† (1.13)	8.21 (.73)	7.48† (.73)
Obtains history	8.38 (1.12)	8.15‡ (.66)	7.70‡ (.60)
Physical exam	8.39† (0.99)	8.20 (.67)	7.62† (.54)
Orders tests appropriately	8.44 (0.95)	8.12 (.66)	NA
Performs procedures carefully	8.29† (1.66)	8.15‡ (.76)	7.69†‡ (.61)
Demonstrates integrity	8.73 (0.79)	8.07 (.55)	8.18 (.32)
Understands role of team	8.39 (1.24)	8.11 (.62)	NA
Responsive, cooperative	8.61 (0.92)	8.13 (.61)	NA
Clinical judgement	8.27 (1.20)	8.10 (.76)	7.59 (.77)
Overall rating	8.39 (0.97)	8.11 (.67)	7.63 (.77)

**Nine-point global rating scale; intern is the unit of analysis. NA indicates not applicable.*

†*Significance between resident rating and faculty rating of interns <.05.*

‡*Significance between intern rating and faculty rating of interns <.05.*

Table 2. Factor Analysis of Peer Evaluations of Interns*

	Senior Resident Evaluators		Intern Evaluators	
Factors	**1. Technical**	**2. Interpersonal**	**1. Technical**	**2. Interpersonal**
Medical knowledge	.94		.88	
History taking	.89		.76	.59
Physical exam	.77		.81	
Orders tests appropriately	.93		.82	
Procedures		.62	.85	
Integrity		.87		.92
Teamwork		.93		.84
Cooperative		.85		.92
Judgment	.98		.81	
Overall	.96		.63	.72
Variance accounted for, %	54.7	29.7	47.4	41.8

**Factor analysis, determined by Varimax factor rotation, is used to identify groups of items in the evalvation instrument which receive similar ratings by evaluators. Numbers reported are factor loadings, which are an index by which an item is associated with a given factor. Factor loadings of greater than .55 are reported. The percent of variance accounted for indicates the extent to which the factor accounted for all of the ratings received.*

resident and intern assessments of intern clinical judgment and overall competence, and between faculty and senior resident assessments of humanistic and professional behaviors of interns.

Although the average number of instruments returned per intern in this study did achieve the number previously shown to be reliable for practicing physicians,[9–11] the low response rate, particularly by interns, introduces the risk of sampling error. Residents gave several reasons for resistance to completing peer forms: paperwork burden, which would have particularly affected interns; lack of clarity in the form itself; and concern that the process would undermine the team function. We suspect that discomfort with the feedback process was an unspoken barrier for many house officers, who had no formalized training in feedback or evaluation. Others have also found marked resistance to the use of peer review, with senior residents being more accepting.[14]

The differential response rate between the two firms was an unexpected finding. Although the means between firms were not significantly different, the peer evaluation did identify two interns in one firm who were performing below average for the firm. Whether the impact of these two interns was sufficient to diminish the response rate overall within the firm is not clear. If so, peer review may indicate the collegial health of the firm as well as provide individual feedback. Further studies over time and in other programs may clarify this issue.

How can the use of peer review be advanced in training programs? We suggest that program directors draw from the experience in the introduction of self-assessment into a number of health professions' curricula.[17] Successful curricula have recognized the need for a transition period that may be characterized by hostility and resistance, and have addressed resident concerns by including residents in the planning body of the evaluation system, by explicit rules concerning confidentiality and process of information gathering, and by additional training in the skills of feedback. Engebretsen's successful model of peer review in a residency system incorporated many of these elements.[13] Unless they have had experience with peer review in medical school settings, it is unlikely that interns,

Table 3. Correlation of Faculty, Senior Resident, and Intern Evaluations of Interns by Instrument Item*

Item in Peer/Faculty Instrument	Faculty and Senior Resident Evaluations	Senior Resident and Intern Evaluations	Faculty and Intern Evaluations
Medical knowledge	.72†	.30	.15
History-taking skills	.60†	.30	.64†
Physical exam	.51	.38	.60‡
Procedural skills	.60	.73†	.52‡
Integrity, compassion/humanism	.31	.44	.57‡
Integrity, compassion/professionalism	.17	—	.49
Clinical judgment	.66†	.19	.60‡
Overall competence	.61†	.16	.50‡

**Pearson product-moment correlation; intern is the unit of analysis.*
†$p < .01$.
‡$p < .05$.

the most vulnerable learners, will be able to quickly adopt peer review, and one approach may be to use senior residents exclusively as "peer" evaluators. We anticipate that the process of specific training in evaluation and the completion of the peer instruments will require residents and faculty to mutually define the meaning of integrity, teamwork, and cooperation, and allow opportunities to bring these competencies of professionalism to the forefront of the training program agenda.

The authors thank Elizabeth Garrett for statistical review and support, and John Shatzer for critical review of earlier versions of this manuscript.

REFERENCES

1. Arnold RM, Povar GJ, Howell JD. The humanities, humanistic behavior, and the humane physician: a cautionary note. Ann Intern Med. 1987;106:313–8.
2. Merkel WT, Margolis RB, Smith RC. Teaching humanistic and psychosocial aspects of care: current practices and attitudes. J Gen Intern Med. 1990;5:34–41.
3. Branch WE, Arky RA, Woo B, Stoeckle JD, Levy DB, Taylor WC. Teaching medicine as a human experience: a patient-doctor relationship course for faculty and first-year medical students. Ann Intern Med. 1991;114:482–9.
4. Woolliscroft JO, Howell JD, Patel BP, Swanson DB. 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.
5. DiMatteo MR, DiNicola DD. Sources of assessment of physician performance: a study of comparative reliability and patterns of intercorrelation. Med Care. 1981;19:829–39.
6. Gough HG, Hall WB, Harris RE. Evaluation of performance in medical training. J Med Educ. 1964;39:679–92.
7. Kegel-Flom P. Predicting supervisor, peer and self ratings of intern performance. J Med Educ. 1975;50:812–5.
8. Arnold L, Willouby L, Calkins V, Gammon L, Eberhardt G. Use of peer evaluation in the assessment of medical students. J Med Educ. 1981;56:35–42.
9. Ramsey PG, Carline JD, Blank LL, Wenrich MJ. Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. Acad Med. 1996;71:364–70.
10. Violato C, Marini A, Towes 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 1):S82–4.
11. 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.
12. Johanson WG. The ABIM recertification program—nearing liftoff. Chest. 1995;108:1–2.
13. Engebretsen B. Peer review in graduate education. N Engl J Med. 1977;296:1230–1.
14. Van Rosendaal GMA, Jennett PA. Resistance to peer evaluation in an internal medicine residency. Acad Med. 1992;67:63.
15. Holmboe ES, Hawkins RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. Ann Intern Med. 1998;129:42–8.
16. American Board of Internal Medicine. Guide to the Evaluation of Residents in Internal Medicine 1990–2. Philadelphia, Pa: American Board of Internal Medicine; 1992.
17. Gordon MJ. Self-assessment programs and their implications for health professions training. Acad Med. 1992;67:62–9.

APPENDIX A

Peer Review Evaluation Form — Inpatient Service

Please evaluate the house officer's performance for each component of clinical competence. Circle the rating which best describes the house officer's skills and abilities. Use your standard level of skill expected from the clearly satisfactory house officer at this stage of training. Identify strengths and weaknesses you have observed. For any component that needs attention or you are unable to judge due to insufficient contact with the house officer, please check the appropriate category. Be as specific as possible, including reports of critical incidents in your comments. Global adjectives or remarks such as "good house officer," do not provide a meaningful feedback to the house officer as specific comments.

Superior: far exceeds reasonable expectations; Satisfactory: always meets reasonable expectations and occasionally exceeds; Unsatisfactory: consistently falls short of reasonable expectations.

	Unsatisfactory			Satisfactory			Superior		
1. Medical knowledge	1	2	3	4	5	6	7	8	9
2. Obtains history completely and carefully	1	2	3	4	5	6	7	8	9
3. Performs physical exam accurately and completely	1	2	3	4	5	6	7	8	9
4. Orders tests appropriately	1	2	3	4	5	6	7	8	9
5. Performs procedures carefully and minimizes risk to patients	1	2	3	4	5	6	7	8	9
6. Demonstrates integrity, empathy, and compassion for the patient	1	2	3	4	5	6	7	8	9
7. Understands and appreciates the role of team members	1	2	3	4	5	6	7	8	9
8. Responsive, cooperative, respectful, timely	1	2	3	4	5	6	7	8	9
9. Clinical judgment: puts together the whole picture	1	2	3	4	5	6	7	8	9
10. Overall rating	1	2	3	4	5	6	7	8	9

Comments: