ORIGINAL ARTICLES

Effectiveness of a Focused Educational Intervention on Resident Evaluations from Faculty

A Randomized Controlled Trial

Eric S. Holmboe, MD, Nicholas H. Fiebach, MD, MPH, Leslie A. Galaty, MAT, Stephen Huot, MD, PhD

OBJECTIVE: To improve the quality and specificity of written evaluations by faculty attendings of internal medicine residents during inpatient rotations.

DESIGN: Prospective randomized controlled trial.

SETTING: Four hospitals: tertiary care university hospital, Veterans' Administration hospital, and two community hospitals.

PARTICIPANTS: Eighty-eight faculty and 157 residents from categorical and primary-care internal medicine residency training programs rotating on inpatient general medicine teams.

INTERVENTION: Focused 20-minute educational session on evaluation and feedback, accompanied by 3 by 5 reminder card and diary, given to faculty at the start of their attending month.

MEASUREMENTS AND MAIN RESULTS: Primary outcomes: 1) number of written comments from faculty specific to unique, preselected dimensions of competence; 2) number of written comments from faculty describing a specific resident behavior or providing a recommendation; and 3) resident Likert-scale ratings of the quantity and effect of feedback received from faculty. Faculty in the intervention group provided more written comments specific to defined dimensions of competence, a median of three comments per evaluation form versus two in the control group, but when adjusted for clustering by faculty, the difference was not statistically significant ($P = .09$). Regarding feedback, residents in the intervention group rated the quantity significantly higher ($P = .04$) and were significantly more likely to make changes

in clinical management of patients than residents in the control group ($P = .04$).

CONCLUSIONS: A brief, focused educational intervention delivered to faculty prior to the start of a ward rotation appears to have a modest effect on faculty behavior for written evaluations and promoted higher quality feedback given to house staff.

KEY WORDS: written evaluation; residents; faculty; educational intervention; controlled trial.

J GEN INTERN MED 2001;16:427-434.

Program directors and attending physicians have a major responsibility to evaluate the clinical competencies of residents and to provide them with timely, useful feedback to ensure continued progress and correct deficiencies.¹⁻³ Program directors are being held increasingly accountable for the performance of their house staff.⁴ All residency programs require faculty to complete an evaluation of residents at the end of a clinical rotation. The most commonly used form is the longitudinal rating scale evaluation form developed by the American Board of Internal Medicine that evaluates components such as clinical judgment, medical knowledge, clinical skills (history-taking, physical examination), humanism, professionalism, and overall medical care.^{5,6} Performance for each component is rated on a 9-point scale: 1-3 denotes unsatisfactory performance, 4 marginal, 5-6 satisfactory, and 7-9 superior. In addition, space is provided on the form for written comments.

Although the rating scale form was designed to be easy to use and to facilitate longitudinal evaluation of a broad array of clinical skills, studies suggest that the use of the rating scale evaluations have several serious limitations.⁷⁻⁹ Two studies found that ratings from attendings failed to discriminate between performance of the various components of competence because of halo effect and range restriction errors.^{8,9} In one study, 96% of over 1,000 ratings were between 6 and 9.⁸ Another study suggests training can reduce halo error in numeric ratings, but this may occur at the expense of accuracy.¹⁰

Received from the Yale University Primary Care Internal Medicine Residency Program, New Haven, Conn (ESH, LAG, SH); and Johns Hopkins General Internal Medicine Residency Program, Johns Hopkins Bayview Medical Center, Baltimore, MD (NHF).

Portions of this work were previously presented at the 22nd Annual Meeting of the Society of General Internal Medicine, San Francisco, Calif, April 29-May 1, 1999.

Address correspondence and reprint requests to Dr. Holmboe: Director, Faculty Development and Evaluation, Department of Internal Medicine, Yale University School of Medicine, 1074 LMP, 333 Cedar St., New Haven, CT 06510 (e-mail: eholmboe@msn.com).

Given the problems associated with the numeric ratings assigned by attendings to the components of competence, written comments are necessary to provide more specific information about a resident's performance. However, the experience of the authors is that most written comments are global and non-specific in nature. Furthermore, few studies have investigated whether faculty training can improve the quality of written evaluations provided on the rating scale forms.⁷

The primary objective of this study was to investigate the effectiveness of a brief educational intervention with faculty to increase the quantity and improve the specificity of written comments on the ABIM evaluation form. We hypothesized that a brief educational intervention would lead to an increase in the quantity of written comments that were specific to a particular domain of competence (e.g., clinical judgment, history-taking skills, humanism, etc.) and that provided either a behavioral example or recommendation. An additional objective was to assess whether the intervention led to more feedback as perceived by the residents.^{11,12}

METHODS

Subjects and Setting

Eighty-eight faculty members and 157 residents (postgraduate years [PGY] 1–3) of the Yale University Department of Medicine assigned to inpatient general internal medicine rotations during academic year 1997–1998 were asked to participate in the study. Faculty was randomized to the intervention or control group only once for the academic year during their first general medicine ward rotation. Subsequent rotations for faculty attending more than once during the academic year were not included. Residents of randomized faculty were eligible throughout the year and could participate more than once. Residents were not informed whether their attending was randomized to the intervention or control group. Residents who spent less than two weeks with the study attending were excluded. Four inpatient settings were included: a university hospital, a Veterans' Administration hospital, and two community-based hospitals.

Evaluation Form

A standard evaluation form based on the rating scale developed by the American Board of Internal Medicine (ABIM) was used for this study.⁶ The form was the same as that used during the previous academic year, 1996–1997. Attendings rated residents, using a 9-point scale, on the dimensions of clinical judgment, medical knowledge, clinical skills, humanistic qualities, teaching ability, professional behavior, medical care, and overall clinical competence. The form explicitly states that a written statement is required with space available at the bottom of the form for comments.

Study Design and Intervention

The study was a prospective randomized controlled trial. The unit of randomization was the attending, and to minimize contamination, randomization of attendings was stratified by hospital prior to each rotation.

Both the control and intervention group attendings received an orientation packet at the beginning of the rotation that was standardized across all four study sites. The standard packet contained an outline of teaching responsibilities, expectations that feedback should be provided to the residents formally during the rotation, and outlined administrative responsibilities regarding notes and dictations. A monograph on effective clinical teaching was also included.¹³

The intervention consisted of a brief 20-minute didactic session provided to attending faculty by one of the 3 investigators (EH, SH, NF) outlining characteristics of effective evaluation and feedback,^{11,12,14} and stressing the importance of direct observation to evaluate clinical competence. The majority of attendings (41 of 47) received the intervention one-on-one. Each investigator used a standard script to deliver the didactic session. Each attending also received a personalized 5" by 6" card for each of their residents. The card was folded to create a 3" by 5" mini-booklet with the front cover of the card providing reminders about the dimensions of competence and key points regarding effective evaluation and feedback (Appendix). Ample space on the remainder of the pocket card was left blank to allow the attending to record comments and observations made during the course of the rotation. Attendings were informed that the cards would not be analyzed and that any observations written on the cards should be transferred to the evaluation form. Attendings were directly encouraged to provide specific written comments on the evaluation form. To minimize bias from the Hawthorne effect in the intervention group,¹⁵ control group attendings were also encouraged to provide specific written comments on their evaluation forms of the residents and told that the investigators would be analyzing these comments. The human investigation committee at all 4 sites approved the protocol.

Creation of Taxonomy

A taxonomy of the written comments was developed prior to initiation of the trial to enable a structured analysis of the written evaluations. To protect resident anonymity, all identifying information, including PGY year, gender, and names, was removed and the comments were transcribed into a document for analysis. Resident evaluations were identified only by a code number. One investigator first analyzed thirty evaluation forms from the 1996–1997 academic year (EH) to produce a preliminary taxonomy, which was then refined by consensus among the three investigators (EH, SH, NF). Each investigator then analyzed a second set of 30

evaluations independently using the preliminary taxonomy. Results among the three investigators were compared to reach consensus and make additional refinements to finalize the taxonomy. To assess inter-rater reliability of the final taxonomy, a third set of written comments from 30 evaluation forms were independently categorized by each of the three investigators using the final refined taxonomy.

Comments from the randomized trial were then categorized, using this taxonomy, by an investigator (EH) blinded to the attending, resident, and randomization status. For the written comments, no information about resident name, gender, or PGY year was transcribed.

Resident Survey

At the end of each rotation, residents completed a questionnaire regarding verbal feedback and observation by their attending. The survey specifically asked 1) if they received any feedback during the rotation, 2) if they received any formal feedback at the midpoint of the rotation, and 3) whether feedback from their attending led to changes in reading habits or how they managed clinical problems. Residents were also asked whether the attending directly observed them performing a history or physical examination. Finally, residents were asked to rate the overall quantity of the feedback they received during the course of the rotation on a 7 point scale, where 1–3 represented too little feedback, 4 was just the right amount of feedback, and 5–7 represented too much feedback.

Statistical Analysis

Demographics, scores from the rating scale, and the taxonomic classifications were entered into an Excel database (Microsoft Corp., Redmond, Wash). Inter-rater agreement between the investigators for the taxonomy of written comments was measured by percent agreement and the κ statistic. Based on an analysis of 30 random evaluations from the 1996–1997 academic year, six percent of the forms contained a behavioral or example comment. Using this baseline, the study had 80% power to detect a 100% increase in the number of behavioral or example comments at an $\alpha = 0.05$.

Differences in demographic characteristics between the two groups were assessed with the χ^2 test. Differences between rating scale scores for each dimension of competency on the evaluation form were measured with the Wilcoxon rank sum test. Using the resident as the unit of analysis, differences between the intervention and control groups regarding the number of dimension-specific and example/behavioral comments were first compared with the Wilcoxon rank sum test. Given that the faculty, however, were the unit of randomization, a Huber/White/sandwich (HWs) regression model with robust standard errors was used to adjust for clustering by faculty for the

written comments and resident questionnaires.^{16,17} The regression model analyses were performed with STATA software (version 5.0, College Station, Tex). All other statistical analyses were performed using SAS software (version 6.12, Cary, NC).

RESULTS

Characteristics of attendings and residents are shown in Table 1. Because the protocol had not been approved by one of the four institutions in time for the first rotation of the academic year, and because these attendings had been randomized to the control group, there are more attendings in the intervention group. Distribution of faculty by academic rank, location of attending assignment, and number of months attending was similar between the two groups. Although there were more women attendings in the intervention group, the difference was not statistically different and the number of evaluations performed by women faculty in the control and intervention groups was not statistically different either ($P = .17$).

The total number of residents who participated in the study was 157, with 42 in the PGY-3 class, 37 in PGY-2, and 78 PGY-1 interns. Table 1B provides a comparison of the housestaff characteristics between the two groups. Because residents could be randomized more than once, the combined total number of residents from the intervention and control groups is greater than absolute number of residents in the trial. There were no statistically significant differences between the two groups with regards to number of women residents ($P = .94$) or PGY year ($P = .38$).

Taxonomy

The taxonomy developed for the written comments is shown in Table 2. There are four categories for the

Table 1A. Attending Demographics

	Intervention Group % (n)	Control Group % (n)	P Value*
Total participants	53 (47)	47 (41)	
Women	26 (12)	15 (6)	.21
Faculty rank			.81
Professor	17 (8)	15 (6)	
Associate Professor	32 (15)	32 (13)	
Assistant Professor	26 (12)	34 (14)	
Instructor	26 (12)	19 (8)	
Location			.69
University hospital	38 (18)	49 (20)	
Veterans' hospital	30 (14)	29 (12)	
Community hospital A	13 (6)	7 (3)	
Community hospital B	19 (9)	15 (6)	
Attended two or months	30 (14)	27 (11)	.75

* Calculated with the χ^2 test.

Table 1B. Resident Demographics

	Intervention Group N = 109	Control Group N = 87	P Value*
% of residents (n)			.47
PGY-3	22 (24)	30 (26)	
PGY-2	24 (26)	20 (17)	
PGY-1	54 (59)	62 (54)	
Number of evaluations	144	126	.38
% PGY-3 (n)	22 (31)	26 (33)	
% PGY-2 (n)	23 (33)	17 (21)	
% PGY-1 (n)	55 (80)	57 (72)	
% women (n)	40 (44)	40 (39)	.94

* Calculated with the χ^2 test.

written comments: global, dimension specific, example or behavioral, and recommendations. The subcategories in the dimension specific category were clinical judgment, medical knowledge, clinical skills, humanism, teaching ability, professionalism, and medical care, leadership, self-directed learning, and organizational skills. Inter-rater agreement for the taxonomy using the final evaluation set was excellent. Between investigators EH and SH, the percent agreement was 84% with a $\kappa = 0.75$, and between EH and NF the percent agreement was 87% with a $\kappa = 0.79$.

Written Evaluations

For the main outcome of this study, residents in the intervention group received a median of three dimension specific comments per evaluation form compared to a median of two comments per evaluation form in the control group ($P = .04$) (Table 2). However, when adjusted for clustering by faculty, the difference did not reach statistical significance ($F = 2.9$; HWs coefficient, 0.81 [95% confidence interval (CI), -13 to 1.75]). More residents in the intervention group received a behavioral/example comment (20%) compared to the control group residents (12%), but the difference was not statistically significant ($F = 0.92$; HWs coefficient, 0.06 [95% CI = -0.07 to 0.20]).

Rating Scale Scores

As shown in Table 3, intervention group residents received slightly lower average scores for the 7 specific dimensions of competence and overall competence, but none of these small differences were meaningful or statistically significant.

Resident Questionnaire

Table 4 provides results from the resident survey. The response rate for the survey was 86%. Forty-four percent of residents in the intervention group, compared with 28% of

Table 2. Taxonomy of Written Comments

Category	Total Number of Comments		Examples
	Intervention (N = 138)* n (%)	Control (N = 120)* n (%)	
Global	280 (33)	242 (38)	a. "Very good resident." b. "(Name) is a solid resident."
Dimension-specific	512 (60)	348 (54)	a. "(Name)'s knowledge is only average." — <i>Medical knowledge</i> b. "(Name) showed great compassion for patients." — <i>Humanism</i> c. "Noted to have problems assessing patients and formulating management plan." — <i>Clinical judgment</i>
Median (mean)			
dimension-specific/form	3 (3.2) ^{†,‡}	2 (2.4) ^{†,‡}	
Example/behavioral	27 (3)	14 (2)	a. "Thorough exam — documented new subtle neuro findings on patient with brain metastases." b. "Made outstanding diagnosis of new hyperparathyroidism in an elderly man."
Recommendation	38 (4)	35 (5)	a. "Needs greater attention to detail on care of complex patients." b. "Now (name) should focus on increasing fund of knowledge."
Total comments	857	639	

* Number of evaluation forms per group.

[†] $P = .04$; Wilcoxon rank sum, resident as unit of analysis.

[‡] $P = .09$; Huber/White/sandwich regression method, adjusting for clustering by faculty.

Table 3. Evaluation Form Ratings

Dimension	Intervention			Control			P Value*
	Mean ($\pm$ SD)	Median	Range	Mean ($\pm$ SD)	Median	Range	
Clinical judgment	6.9 (1.3)	7	3 to 9	7.2 (1.3)	7	4 to 9	.11
Medical knowledge	6.8 (1.3)	7	4 to 9	7.0 (1.2)	7	4 to 9	.23
Clinical skills	7.0 (1.2)	7	4 to 9	7.3 (1.1)	7	4 to 9	.11
Humanism	7.7 (1.1)	8	4 to 9	7.7 (1.1)	8	4 to 9	.75
Teaching	7.2 (1.4)	7	3 to 9	7.4 (1.2)	8	4 to 9	.24
Professionalism	7.8 (1.1)	8	4 to 9	7.8 (1.1)	8	5 to 9	.73
Medical care	7.5 (1.1)	8	4 to 9	7.6 (1.2)	8	4 to 9	.56
Overall competence	7.2 (1.2)	7	4 to 9	7.4 (1.1)	8	4 to 9	.26

* P value calculated with Wilcoxon rank sum test.

the control group, stated that feedback from their attending led them to make changes in their approach to the management of clinical problems (adjusted for clustering by faculty, $F = 4.22$; HWs coefficient, 0.15 [95% CI, 0.005 to 0.30]; $P = .04$). Regarding the quantity of overall feedback, residents in the intervention group provided a median rating of 3.5 versus 3 for the control group, with a rating of 4 representing "just the right amount" of feedback and less than four representing too little feedback ($P = .03$). After adjustment for clustering by faculty, this difference also remained statistically significant ($F = 4.21$; HWs coefficient, 0.43 [95% CI = 0.01 to 0.84]; $P = .04$). There were no statistically significant differences between the two groups for the number of residents receiving feedback during the rotation, receiving formal midrotation feedback, or being observed performing part of a physical exam or history.

DISCUSSION

This study demonstrates that a simple, brief intervention with attending faculty produced a modest change in the written evaluations of residents, but had a larger effect on the feedback to residents on an inpatient general medical service. The intervention group faculty provided more dimension-specific comments and behavioral examples than faculty in the control group, but only the number of dimension-specific comments approached statistical significance when adjusted for the faculty as the unit of analysis. These findings have important implications for resident evaluation. It is well documented that faculty

provide a paucity of written comments on evaluation forms,^{4,7,8} even when specific resident behaviors are observed.⁴ Gray recently noted that "the lack of written comments on many carefully studied rating scales such as the ABIM form creates problems for the program director who wished to use the form to provide feedback."⁷ Given that numeric ratings from attendings often do not discriminate between the various dimensions of clinical competence,^{8,9,18} written comments are vital to provide insight into a resident's specific strengths and weaknesses.

This is among the first studies to rigorously examine the effects of an intervention on written evaluations. Given the importance of written comments to residency program directors, interventions to promote more specific faculty evaluations are needed.⁶ The statistical power of our study only allows us to suggest that a brief educational intervention may improve the specificity and nature of the written comments. However, from the viewpoint of "clinical" significance, an increase of just one specific comment per rotation extrapolates to 30–40 additional specific comments over the course of a residency. This finding may be particularly pertinent to residency programs that do not have the resources to utilize other, more expensive evaluation tools such as standardized patients and computer simulations.¹⁸ More work is needed to corroborate our findings as well as investigate whether repeated, brief interventions would be more effective. This study also raises the sobering possibility that more intensive interventions will be needed to change faculty behavior, a very difficult task for busy clinician-educators experiencing significant time

Table 4. Results of Resident Survey

Question	Intervention Group	Control Group	P Value*
Residents receiving formal feedback at midpoint of rotation, %	30	23	.34
Residents who made changes in clinical management based on attending's feedback, %	44	28	.04
Residents observed performing some aspect of physical exam, %	42	45	.59
Residents observed obtaining some portion of clinical history, %	13	16	.69
Overall rating of quantity of feedback from attending (median)	3.5	3	.04

* P value calculated with Huber/White/sandwich regression method adjusting for clustering by faculty.

constraints on teaching activities. Despite the modest effect on written faculty evaluations, the intervention did lead to important changes at the resident level. Residents in the intervention group were more likely to report that they received an appropriate or “right” amount of feedback. Perhaps more importantly, the finding that residents in the intervention group reported that feedback from faculty was more likely to change their approach to patient management has potentially important implications. This clinically significant observation, if confirmed, suggests that this educational intervention can positively influence the care of patients and therefore affect the quality of care delivered in teaching hospitals.

We were also able to demonstrate that written qualitative comments can be reliably categorized into a working taxonomy. This taxonomy allowed for a rigorous analysis of the written comments, the primary outcome of this study. This simple taxonomy, based on the dimensions of competence already defined by the ABIM and other groups, can help program directors to evaluate the quality and specificity of their program’s own written evaluations.

Another important strength of this study was the prospective randomized design that allowed for a rigorous evaluation of the effect of the intervention, including appropriate use of blinding in the evaluation of outcomes and assessment of inter-rater reliability for the analysis. It is very important to not only develop practical strategies to improve evaluation but it is equally important to also rigorously evaluate their effectiveness. Recently, several authors have specifically highlighted the need for more rigorous methodology to study educational interventions.^{19,20} Given the growing public concern over the competence of graduating trainees, training programs must work with faculty to ensure they perform more detailed evaluations of residents’ knowledge, skills, and attitudes.

The study was also designed to minimize the impact of the Hawthorne effect on the results of the intervention.¹⁵ Faculty in the control group were clearly aware that they were involved in an evaluation study and were informed that their written comments would be analyzed. Control group attendings were also encouraged to provide feedback to residents during the rotation. Both intervention and control group attendings may have changed their behaviors as a consequence of participating.

This study has several limitations. First, the impact of the intervention was modest and not all aspects of the evaluation process demonstrated improvement. Our sample was ultimately not large enough to detect small but significant differences between the groups, especially using the attending as the unit of analysis. Control group attendings also showed substantial proportional improvement, most likely as a result of the Hawthorne effect and the observation that baseline performance was so poor. This may have narrowed the differences between the intervention and control groups, so that the results may underestimate the true impact of the intervention.

Despite the lack of statistical significance, the intervention group did provide three-fold more behavioral or example comments over baseline, and 67% more than the control group.

Second, residents in the intervention group did not report more formal feedback sessions at the mid-point of their rotation or greater amounts of direct observation of history and physical examinations skills from their attendings. However, the educational intervention was not designed to specifically target these areas as a primary outcome. Third, mean numeric ratings on the evaluation forms were not significantly different between the two groups, but this was not a designed outcome of the study and is consistent with previous reports.^{8,9} This lack of observed difference in numeric ratings actually reinforces the importance of improving the quality of the written comments on the evaluation form. Several authors have pointed out the need to improve the quality of written comments precisely because of the failure of numeric ratings to discriminate between dimensions of clinical competence.^{4,7} Finally, the numeric ratings suggests that the actual, observed average competencies were similar between the two groups and the intervention group did not “deserve” more specific comments because they had more problems.

The study was also not designed to investigate the reliability or validity of the attendings’ written evaluations. Previous studies of clinical evaluation interventions have found variable reliability, even of the numeric rating.^{8,9,21–24} Although rating scale evaluations of residents have been shown to have modest predictive validity for the ABIM certification exam scores,²⁵ the accuracy and other aspects of the validity of rating scales has not been substantiated. Problems with the accuracy of attending evaluations when compared with other evaluation methods, such the clinical evaluation exercise, have been documented.^{26–30} More work is needed to assess the reliability and validity of written evaluations.

Finally, the educational intervention used in this study was not comprehensive and did not address the issue of sustainability over time. As noted, however, the focused nature of the intervention may be a strength of the study that would allow for repeated interventions with faculty over time to reinforce key behaviors in the evaluation of clinical competence. This approach deserves further study. Additionally, the intervention is practical, and when combined with other tools to measure competence such as the mini-CEX, may help to create a more comprehensive evaluation of residents over time.²⁹ More research is needed to examine the effect of more comprehensive faculty development interventions as well as the sustainability of interventions to improve evaluation.

CONCLUSIONS

A brief, focused educational intervention, accompanied by a practical 3” by 5” reminder card and note diary,

delivered to faculty prior to the start of a ward rotation may help to modestly change faculty behavior in the written evaluation of and feedback given to residents. Given the importance of comprehensive, specific evaluations of residents in an era when attendings are continually pressed for time, this intervention may also provide a simple yet effective method to improve the feedback given to residents. Rigorous research designs to assess effectiveness of educational interventions to improve the evaluation of competence will be critical to defining optimal approaches for resident evaluation by faculty.

The authors would like to thank Yun Wang, MS, for assistance with the statistical analysis, and Patrick O'Connor, MD, MPH, for his review of the manuscript.

Dr. Holmboe was a Robert Wood Johnson Clinical Scholar during completion of this work.

REFERENCES

1. Eisenberg JM. Evaluating internists' clinical competence. *J Gen Intern Med.* 1989;4:139-43.
2. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med.* 1990;65(suppl):63-7.
3. American Board of Internal Medicine. Project Professionalism. Philadelphia: American Board of Internal Medicine; 1994.
4. Turnbull J, Gray J, MacFadyen J. Improving in-training evaluation programs. *J Gen Intern Med.* 1998;13:317-23.
5. American Board of Internal Medicine. Attending Physicians: Your Role in Evaluating Residents. Philadelphia: American Board of Internal Medicine; 1995.
6. American Board of Internal Medicine. Guide to the Evaluation of Residents in Internal Medicine 1990-92. Philadelphia: American Board of Internal Medicine; 1992.
7. Gray JD. Global rating scales in residency education. *Acad Med.* 1996;71:S55-63.
8. Thompson WG, Lipkin M, Gilbert DA, Guzzo RA, Roberson L. Evaluating evaluation: assessment of the American Board of Internal Medicine resident evaluation form. *J Gen Intern Med.* 1990;5:214-7.
9. Haber RJ, Avins AL. Do ratings on the American Board of Internal Medicine resident evaluation form detect differences in clinical competence? *J Gen Intern Med.* 1994;9:140-5.
10. Borman WC. Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *J Appl Psychol.* 1975;60:556-60.
11. Ende J. Feedback in clinical medical education. *JAMA.* 1983;250:777-81.
12. Ende J. The evaluation product: putting it to use. In: Lloyd JS, Langley DG, eds. *How to Evaluate Residents.* Chicago: American Board of Medical Specialties; 1986:99-116.
13. Weinholtz D. Teaching during attending rounds: a manual for attending physicians. Des Moines, Iowa: University of Iowa; 1987.
14. Skeff K, Stratos G. Feedback. From the Stanford Faculty Development Program, Clinical Teaching. Stanford, Calif: Stanford University; 1996.
15. Neufeld V, Norman GR, eds. *Assessing Clinical Competence.* New York: Springer; 1985.
16. Huber PJ. The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium in Mathematical Probability.* Berkeley, Calif: University of California Press; 1967.
17. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica.* 1980;48:817-30.
18. Holmboe ES, Hawkins RE. Methods for evaluating the clinical competence of residents in Internal Medicine: a review. *Ann Intern Med.* 1998;129:42-8.
19. Veet LL, Shea JA, Ende J. Our continuing interest in manuscripts about education. *J Gen Intern Med.* 1997;12:583-5.
20. Education Group for Guidelines on Evaluation. Guideline for evaluating papers on educational interventions. *BMJ.* 1999;318:1265-7.
21. Davis JK, Inamdar S, Stone RK. Interrater agreement and predictive validity of faculty ratings of pediatric residents. *J Med Educ.* 1986;61:901-5.
22. Maxim BR, Dielman TE. Dimensionality, internal consistency, and interrater reliability of clinical performance ratings. *Med Educ.* 1996;21:130-7.
23. Carline JD, Paaauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. *J Gen Intern Med.* 1992;7:506-10.
24. Norcini JJ. Standards and reliability in evaluation: when rules of thumb don't apply. *Acad Med.* 1999;74:1088-90.
25. Kassirer JP. Performance results on the 1995 Internal Medicine certifying examination. *ABIM News Update.* 1995;Fall/Winter:2-4.
26. 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.
27. Herbers JE, Noel GL, Cooper GS, Harvey J, Pangaro LN, Weaver MJ. How accurate are faculty evaluations of clinical competence? *J Gen Intern Med.* 1989;4:202-8.
28. Kroboth FJ, Hanusa BH, Parker S, et al. The inter-rater reliability and internal consistency of a clinical examination exercise. *J Gen Intern Med.* 1992;7:174-9.
29. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med.* 1995;123:795-9.
30. Kalet A, Earp J, Kowlowitz V. How well do the faculty evaluate the interviewing skills of medical students? *J Gen Intern Med.* 1992;7:499-505.

APPENDIX

3 by 5 Reminder/Diary Card

Resident: _____

OBSERVATIONS / NOTES

ABIM COMPETENCIES**The 9 Areas for Evaluation**

Clinical judgment	Medical knowledge
History taking	Physical examination
Procedural skills	Humanistic skills
Medical care	Teaching ability
Professional attitudes and behavior	

EVALUATION**Purpose**

Process of assessing learner's knowledge, skills, and attitudes.

Key Characteristics

Record SPECIFIC behaviors as they occur. Use this card as an aide-de-memoir; the attending's diary for the rotation.

REMEMBER: DIRECT OBSERVATION

is critical to effective evaluation and feedback.

FEEDBACK (FB)**Purpose**

Teacher provides learner with information about performance to help learner improve.

Key Characteristics

Be SPECIFIC / Give FREQUENTLY / TIMING key / Provide both REINFORCING and CORRECTIVE FB.
 Allow learner to REACT to feedback.
 Provide BEHAVIORAL feedback.
 Work with resident on ACTION PLAN/GOAL.

REMEMBER: Even great residents need FB!

For poor or superior performance, NOTE specific behaviors that support evaluation.