Are Anonymous Evaluations a Better Assessment of Faculty Teaching Performance? A Comparative Analysis of Open and Anonymous Evaluation Processes

Nelia M. Afonso, MD; Lavoisier J. Cardozo, MD; Oswald A.J. Mascarenhas, PhD; Anil N.F. Aranha, PhD; Chirag Shah, MD

Objective: We compared teaching performance of medical school faculty using anonymous evaluations and open evaluations (in which the evaluator was not anonymous) and examined barriers to open evaluation. Methods: Residents and medical students evaluated faculty using an open evaluation instrument in which their identity was indicated in the evaluation. Following this, they completed anonymous evaluation on the same faculty members. Aggregate outcomes using the two evaluation systems were compared. Outcomes by group of evaluators (residents and students) were analyzed. Trainees were also asked to rate the barriers to the open evaluation process. Results: A statistically significant difference between the open and anonymous evaluations was noted across all items, with faculty receiving lower scores on the anonymous evaluations. The mean score for all the items on the open evaluations was 4.45 ± 0.65, compared to mean score of 4.07 ± 0.80 on the anonymous evaluations. There was also a statistically significant difference between open and anonymous evaluations in five clinical teaching domains that were evaluated individually. Residents perceived that the three most common barriers to optimal evaluation were an apprehension of possible encounters with the same attending physician in the future, destruction of working relationships with the attending, and a feeling of frustration with the evaluation system. Conclusions: The evaluation of faculty teaching performance is complex. Most academic medical centers use the open evaluation format. This study supports the case for the use of the anonymous evaluation method as a more accurate reflection of teaching performance.

(Fam Med 2005;37(1):43-7.)

It is important in academic medical centers for resident physicians and medical students to be able to evaluate faculty teaching performance. The Residency Review Committee for Internal Medicine (RRC-IM), an organization responsible for the accreditation of training programs, requires all program directors to provide opportunities for residents to evaluate their attending physicians. These evaluations provide a basis for faculty feedback, identification of areas for faculty development, prioritizing resource allocation, and assisting in the promotion and tenure process. Indeed, with the emergence of clinician-educator tracks, teaching performance is becoming an increasingly important criterion for faculty advancement. Although many evaluation systems exist, the current evaluation system at most institutions uses an open rather than an anonymous process for trainees (residents and medical students) to evaluate faculty.

In an open evaluation, the evaluator’s identity is known to the faculty member. Despite the widespread use of the open evaluation format, we have a perception that it may tend to inflate faculty performance grades, but there are no studies to support this view. We therefore decided to compare our faculty teaching performance using both open and anonymous evaluation formats. In addition, we analyzed results of two groups of evaluators (medical students and residents) using both systems. Barriers to optimal evaluation with the open process were also examined.

Methods
Study Design
This was a prospective single institution study. It compared the open and the anonymous evaluation of faculty teaching performance by residents and medical students.
Study Population
This study was conducted in the Department of Internal Medicine at a Midwestern medical school. Attending physicians on general internal medicine wards and the coronary care unit were evaluated over a period of 3 months by residents and medical students.

Evaluation Instrument
Our evaluation instrument (Appendix 1) was based on a format that rates the major components of good clinical teaching. It includes a total of 18 items rated on a 5-point Likert scale that address teaching skills, humanistic qualities, medical knowledge, clinical judgment, quality of supervision provided, and the overall effectiveness of the faculty member as a clinical educator.

In addition, we conducted a focus group to identify a list of perceived barriers to optimum evaluation. The focus group was formed from the 10 residents and six students who were on inpatient service the month prior to the study. The trainees were also asked to rate a list of barriers on a 5-point Likert scale and encouraged to add other barriers not listed.

Study Methodology
Residents and medical students evaluated the faculty using the open evaluation instrument. Following this, they were invited to complete an identical anonymous evaluation on the same faculty member. The chief medical resident distributed the evaluations, and, once completed, the trainees placed the anonymous evaluations in a sealed box. This study was not designed to assess differences in individual faculty teaching performance. The intent was rather to compare aggregate outcomes using the two evaluation systems. We also used the data to analyze outcomes by group of evaluators (residents and students).

Statistical Analysis
The data obtained were coded, entered into a data file, and analyzed using SPSS for Windows® software, Release 7.0 (SPSS Inc, 1995). Statistical methods used included reliability testing using Cronbach’s standard coefficient alpha, Pearson’s correlation, ANOVA, and Student’s two sample t test. Results are presented as mean ± standard deviation (SD). Statistical significance for all tests was established at a nominal P level of .05. The two groups were compared, using the Student’s t test. Faculty were not compared on an individual basis. For each of the 18 items on the evaluation, a comparison was made between the composite scores of the open evaluation and the anonymous faculty teaching evaluation system.

Reliability Analysis
To allow for a meaningful conclusion, the evaluation instrument was analyzed for internal consistency and inter-rater reliability. As described in Table 1, reliability coefficients (Cronbach’s alpha) were used to determine the internal consistency for each of the five teaching performance categories.

Based on factor analysis, statements were grouped as shown in Table 1. An analysis of the most frequently perceived barriers (Table 2) to the current evaluation system was also performed. Inter-rater reliability was evaluated using a method described by Love et al. This method makes the assumption that either reverence or hostility toward some of the faculty may distort trainees’ evaluation.

Results
A total of 30 faculty members were evaluated by the residents and students, who completed 199 evaluations. Of these, 88 were anonymous and 111 open. There were 73 open and 43 anonymous evaluations filled out by residents, while students completed 36 open and 21 anonymous evaluations. The source (student or resident) of 26 anonymous evaluations could not be identified, and these were excluded. A total of 83 trainees provided responses about barriers to open evaluation.

Reliability Analysis
The mean reliability coefficient (Cronbach’s alpha) for determining the internal consistency for the five teaching performance categories was 0.9657, indicating a high degree of internal consistency (Table 1). After trainee responses were eliminated two at a time and the data recalculated to evaluate inter-rater reliability, recalculated data for each of the five domains demonstrated a high degree of inter-rater reliability (Cronbach’s alpha values ranged from 0.88–0.95 for resident and 0.89–0.95 for student evaluators).

Differences Between Open and Anonymous Evaluations
An initial analysis was performed using all the evaluations from students and residents. A statistically significant difference (P<.05) between the open and the anonymous evaluations was noted across all 18 items, with the faculty receiving lower scores on the anonymous evaluations. The mean score for all the items on the open evaluations was 4.45 ± 0.65, compared to mean score of 4.07 ± 0.80 on the anonymous evaluations, a highly significant difference (P=.001). When the five domains of clinical teaching were evaluated as separate groups (supervision, medical knowledge, humanistic skills, teaching skills, and overall effectiveness as a clinical teacher), there were also statistically significant differences between the open and anonymous evaluation in all five domains (Table 3). Differences in the evaluation of faculty by the residents as compared to students for the five teaching categories also demonstrated a statistically significant difference between open and anonymous evaluations by residents in all five.
areas (Table 4). In all but one category, students’ anonymous scores were higher, though not significantly.

**Barriers**

The three barriers to optimal evaluation most often cited by residents were an apprehension of possible encounters with the same attending physician in the future, destruction of working relationships with the attending, and a feeling of frustration that the evaluation process would not improve the teaching performance.

Students listed a lack of formal training about evaluating as their primary barrier to effective evaluation. A feeling of frustration with the evaluation process and the fear of possible encounters with the same attending physician in the future were also listed as barriers.

**Discussion**

There have been no prior studies that compared open and anonymous evaluation processes using the same evaluation instrument. The only related research was a study of radiology residents by Pelsang and Smith, which found that anonymous ballots produced a higher rating when compared to in-person debriefing.

We undertook this study, therefore, to test the hypothesis that there are differences in faculty evaluation outcomes between open and anonymous formats. Our instrument was based on the assessment of accepted criteria that contribute to good teaching outcomes, and we were able to show that our evaluation instrument has a high degree of internal consistency.

There is also evidence that learners discriminate accurately between the different aspects of teaching and that a small number of raters are enough to adequately evaluate teaching performance.

There was a statistically significant difference between the two evaluation processes at the resident level, with the open evaluation process producing significantly higher ratings when compared to the anonymous format and with residents rating faculty members higher than students. This is in keeping with prior studies showing that residents consistently rated faculty members higher than students.

A similar phenomenon has also been noted in two studies in the classroom environment, where upper-division graduate students rated teachers more favorably than lower-division students. The lower ranking by the students might reflect a more honest approach on the part of students when evaluating at-
According to the evaluation format used, constructive feedback is necessary, though this is often deficient. We suggest that a method of periodic feedback given to residents by the program director and to the faculty by division chiefs (based on a summation of a number of evaluations over time) would preserve confidentiality while contributing to better outcomes for both the teacher and the learner. It is also important to note that students felt that they were not adequately trained on how to evaluate teaching performance. Unfortunately, education on how to evaluate is not part of the curriculum in most institutions. There is a clear need for this type of training at all levels.

Limitations
While we have made a case for the anonymous evaluation system, it is also important to consider the limitations of this study. First, not all the trainees who rated faculty volunteered to complete the anonymous evaluation. Consequently, the extent to which the ratings would have differed if all the residents and students had been required to complete the anonymous evaluation form cannot be ascertained. It is important that future studies in this area should consider additional methods to preserve anonymity. In addition, this study was conducted at an academic center in a major teaching hospital and may not be applicable to other teaching settings. Moreover, other variables that may influence evaluation, such as academic rank of the attending physician or the level of involvement of the attending with the trainee, were not analyzed. Finally, the limited sample size, while appropriate for our setting, precludes generalization of these results.

Conclusions
The evaluation of faculty teaching performance is a complex process. This study showed statistically significant differences between open and anonymous evaluations. The results support use of the anonymous evaluation method as a more accurate representation of teaching performance. We have highlighted recent evaluation innovations that are Web-based and capable of preserving the anonymity of the evaluator. It is clear, however, that further research is necessary. This must include instruction on how to evaluate and rater education to reduce bias and error, as well as the development of other innovative formats for rating faculty teaching that may help overcome some of the barriers to evaluation. Finally, constructive efforts toward faculty development to enhance teaching performance are also important.
REFERENCES


Appendix 1

Characteristics of Faculty Attending Physician That Were Rated by Trainees

1. Maintains scheduled teaching time
2. Enthusiasm for teaching
3. Rate the clinical supervision provided by the physician for:
   a. Reviewing histories and physicals
   b. Analyzing clinical data and developing diagnostic hypothesis
   c. Suggesting diagnostic studies and procedures
   d. Encouraging you to assume responsibility for patient care
   e. Discussion of information from recent medical literature
   f. Reviewing order writing and record keeping
4. Rate the supervising physician’s bedside teaching skills
5. Rate the helpfulness of feedback that the supervising physician gave you on your performance
6. Rate the supervising physician’s availability for helping you solve day-to-day problems
7. Rate the overall effectiveness of your supervising physician as an educator/supervisor
8. Rate the attending on the basis of clinical excellence:
   a. Are diagnostic treatment plans well thought out?
   b. Is the attending up to date?
   c. Is his/her use of highly technical, invasive, costly tests and procedures appropriate?
   d. Is the attending sympathetic with and considerate toward patients?
   e. Is the attending communicative?
9. What is your assessment of his/her overall degree of clinical excellence?