● STUDENT AND RESIDENT EVALUATIONS OF FACULTY —HOW DEPENDABLE ARE THEY?

Moderator: John Littlefield, PhD

A Comparison of Student Evaluations and Faculty Peer Evaluations of Faculty Lectures

MARTIN H. LEAMON, MARK E. SERVIS, ROBERT D. CANNING, and ROBERT C. SEARLES

The principal instrument of medical student curriculum evaluation has been student ratings of courses and faculty teaching. Student evaluation of teaching (SET) is widely regarded in higher education as "the single most valid source of data on teaching effectiveness."[1] More recently, the importance of faculty peer evaluation (FPE) of teaching has been emphasized.[2] Depending on the evaluation design and content, FPE may bring different skills, perspectives, and expertise to the evaluation process.[3] In general, however, studies of teaching evaluation in university settings have supported a high degree of correlation between SET and FPE, particularly when the latter is accomplished by direct classroom observation.[4–6] Yet there is surprisingly little investigation comparing student evaluations with faculty peer evaluations of teaching in preclinical medical school courses. Preclinical medical school courses may differ significantly from courses in other higher education settings in that medical courses tend to have multiple instructors who may deliver only one lecture each. In the few early studies comparing PFE and SET in such multi-instructor courses, direct peer evaluation either was not mandatory[7] or relied on a single observation by one peer.[8]

Several authors have questioned the validity of FPE when the number of FPEs per evaluee is small.[9,10] Direct peer observation is time-consuming and can be difficult to accomplish in medical school settings, where clinical faculty are under increasing time pressures and where the medical school is not located near the practice sites of most of the clinical faculty. Furthermore, the multi-instructor course, because of its structure, makes direct observation and evaluation of multiple lecturers by adequately large groups of faculty peers impractical. Methods other than direct observation have been used for FPE, such as reviewing course materials, facilitating student feedback groups or collaborative teaching, often depending on whether the FPE is used primarily for faculty development or for personnel promotion purposes.[1,4,11,12] There has, however, been little work examining the validity of FPE in any higher education setting through observation of teaching other than by classroom visitation, such as by video review or by remote viewing. Given the paucity of research in these areas, the current study sought to explore possible differences between medical student evaluations of live and videotaped lectures and faculty peer evaluations of the same videotaped lectures.

In our study comparable numbers of student and faculty peers evaluated two faculty lecturers presenting one lecture each in a second-year medical student course using a multi-dimensional measure derived from the literature. Students evaluated the lecturers in both live and video formats, while faculty peers evaluated the same lecturers in video format only.

Method

Respondents. The medical student evaluators were 40 volunteers from a second-year medical school lecture course of 102 students. The faculty peer evaluators were 31 full- and part-time faculty members of the Department of Psychiatry at the University of California Davis Medical Center in Sacramento, California.

Procedure. Two faculty members in the Department of Psychiatry were recruited to have their lectures videotaped for the study. These faculty members were part of a larger group of 15 faculty who deliver 27 one-hour lectures in the second-year medical school psychiatry course. The lecturers were selected particularly for having had no contact with the class other than the evaluated lectures. Thus, the SETs could not be influenced by students' prior experience with the lecturers. Each lecture was one hour long and used audiovisual and other materials such as slides and syllabus sections. The lectures were videotaped for distribution to student and faculty raters.

Because evaluations of lecturers viewed in person (student raters) and lecturers viewed on videotape (faculty raters) could differ, we divided the medical student raters into two equal groups (n = 20) and randomly assigned one group to the "live condition" and one group to the "video condition" for the first lecturer. The viewing conditions were reversed for the second lecturer.

Student raters in the live condition viewed the lectures as part of their weekly class. Student raters in the video condition were instructed *not* to attend the live lecture but to pick up a videotape of the lecture and printed copies of the slides. They were instructed to view the lecture at their leisure during the same week that their colleagues in the live condition attended the lecture. Faculty raters were recruited either in person or by letter from the authors. Each of the faculty evaluators was given two videotapes (one of each lecturer), rating forms, and course materials identical to those of the student raters.

Measures. Student and faculty raters filled out a course evaluation form immediately after each lecture. The measure and its domains were derived from similar measures used previously to evaluate university teaching.[6,13] The content of the measure focused on three domains of effective teaching: characteristics of the teacher, characteristics of the lecture, and overall effectiveness.[13]

The evaluation form contained 11 items in three sections: teacher characteristics (five items: enthusiasm, ability to stimulate interest, ability to communicate clearly, apparent knowledge of topic, and rapport); characteristics of the lecture and materials (three items: organization, use of audiovisual materials, and provision of handouts); and the overall effectiveness of the lecture (three items: overall importance, quality and amount of information learned). Each question was rated on a scale from 1 (poor/low) to 5 (excellent/high). The three domains had satisfactory internal consistency as measured by Cronbach's alpha (.88, .77, and .86, respectively).

Analysis. Mean scores for the three domains taken together were submitted to a 2 × 2 MANOVA for the student-viewing-condition analysis (video vs live observation and lecturer 1 vs lecturer 2), and a 2 × 2 MANCOVA (faculty vs student evaluators and lecturer 1 vs lecturer 2) for the comparisons of student and faculty evaluators. These analyses were followed by univariate 2 × 2 ANOVA/ANCOVAs of each evaluation domain. Viewing condition (live vs. video) was entered as a covariate in the second set of analyses because of its significant association with the dependent variables.

Results

Student vs Student. The top half of Table 1 shows the mean ratings of live-condition (LC) student ratings and video-condition (VC) student ratings for both lecturers. The multivariate analysis of variance found that the three dimensions of evaluation, when taken together, showed significant differences between lecturer 1

Table 1. Mean Ratings of Teaching Effectiveness Domains by Students Viewing Video or Live Lectures and Faculty Watching Video Lectures, University of California, Davis School of Medicine, 1998

Domain	Lecturer 1		Lecturer 2	
	Ratings of 18 Students, Video Lecture, Mean (SD)	Ratings of 19 Students, Live Lecture, Mean (SD)	Ratings of 15 Students Video Lecture, Mean (SD)	Ratings of 15 Students Live Lecture, Mean (SD)
Student video vs student live*				
Teacher characteristics	3.86 (0.69)	4.07 (0.72)†	3.94 (0.54)	4.41 (0.49)†
Lecture characteristics	4.07 (0.57)‡	3.97 (0.96)‡	3.49 (0.83)	3.88 (0.68)
Overall effectiveness	3.91 (0.67)	3.95 (0.74)†	3.67 (0.54)	4.29 (0.50)†
	31 Faculty Raters, Adj. Mean (SE)	37 Student Raters Adj. Mean (SE)	28 Faculty Raters, Adj. Mean (SE)	30 Student Raters Adj. Mean (SE)
Faculty video vs student live and video§				
Teacher characteristics	3.76 (0.12)	3.89 (0.11)	4.31 (0.13)‖	4.10 (0.12)‖
Lecture characteristics	4.05 (0.15)¶	3.99 (0.13)¶	3.63 (0.15)	3.66 (0.15)
Overall effectiveness	3.86 (0.14)	3.86 (0.13)	4.27 (0.14)**	3.91 (0.14)**

*Analyzed by ANOVA.
†Live viewing > video viewing, $p < .05$.
‡Lecturer 2 > Lecturer 1, $p < .10$.
§Means (standard errors) adjusted for effects of covariate (viewing condition) analyzed by ANCOVA.
‖Lecturer 2 > Lecturer 1, $p < .05$.
¶Lecturer 1 > Lecturer 2, $p < .05$.
**Lecturer 2 > Lecturer 1, $p < .10$.

and lecturer 2 across viewing conditions (Wilks' $\Lambda = .78$, $F(3,61) = 5.67$, $p < .01$), and between the VC and LC student ratings regardless of lecturer (Wilks' $\Lambda = .88$, $F(3,61) = 2.77$, $p < .05$). The univariate analyses showed that for each lecturer, ratings were significantly higher for LC than VC students for Teacher Characteristics ($F(1,63) = 4.90$, $p < .05$) and Overall Effectiveness ($F(1,63) = 4.60$, $p < .05$). There was also a trend for all students to rate lecturer 1's Lecture Characteristics higher than those of lecturer 2, regardless of viewing condition ($F(1,63) = 3.04$, $p < .10$). Finally, there was a mild interaction between viewing condition and lecturer ($F(1,63) = 3.56$, $p < .10$), with VC students' ratings of Overall Effectiveness dropping from lecturer 1 to lecturer 2, while those of LC students increased. Thus, in general, LC students gave higher ratings than VC students on characteristics associated with the individual teaching style of the instructor (Teacher Characteristics) and his or her Overall Effectiveness. Also, regardless of viewing condition, students differentiated between the two lecturers when the domains were taken together.

Faculty vs Students. The bottom half of Table 1 presents the mean ratings (adjusted for the effects of the covariant viewing condition) of student and faculty evaluators for both lecturers. There was no multivariate difference between ratings by faculty and student evaluators when the effects of viewing condition and lecture were controlled. Teacher characteristics, lecture characteristics, and overall effectiveness, when taken together, were rated significantly different across evaluator condition for lecturer 1 and lecturer 2 (Wilks' $\Lambda = .67$, $F(3,119) = 19.84$, $p < .001$). As in the student analysis, the faculty evaluators made a significant distinction between the effectiveness of the two lecturers and their materials. There was also a trend toward significance in the interaction of evaluator and lecturer (Wilks' $\Lambda = .95$, $F(3,119) = 2.31$, $p < .10$), suggesting that evaluations by faculty and students differed based on which lecturer was being rated.

The univariate analyses of the two groups revealed no significant difference overall or on any of the dimensions of teaching effectiveness between faculty evaluators and student evaluators, after controlling for viewing condition and across both lecturers. Second, the analyses showed significant, or near-significant, differences based upon lecturer, with lecturer 2 being rated higher than lecturer 1 on teacher characteristics ($F(1,121) = 11.40$, $p < .05$), Lecturer 1 rated higher than Lecturer 2 on lecture characteristics ($F(1,121) = 7.61$, $p < .05$), and, finally, a tendency for lecturer 2 to be rated higher on overall effectiveness than lecturer 1 ($F(1,121) = 3.16$, $p < .10$). Finally, the covariate, viewing condition, continued to have a significant effect on ratings of teacher characteristics ($F(1,121) = 4.53$, $p < .05$), and a trend toward significance for overall effectiveness ($F(1,121) = 2.92$, $p < .10$).

Discussion

The study's most salient finding is the absence of differences between student ratings of medical school lecturers' effectiveness and faculty ratings of the same lecturers, after controlling for the effects of viewing condition. Students who viewed lectures in a live lecture hall setting did rate characteristics of the teacher and a lecturer's overall effectiveness higher than did those students who viewed the same lectures on tape. This result may reflect the educational advantage afforded by the personal connection between students and a teacher that a live lecture format allows, compared with the more impersonal relationship of viewing a videotape. It is interesting that the students did not rate the structural/physical characteristics of the lectures (organization, use of audiovisual materials, and syllabus handouts) differently across viewing conditions. The finding that both students and faculty rated the lecturers' effectiveness differently, discriminating between the two lecturers, supports the face validity of the study, as one might expect two lecturers to differ in their effectiveness and thus have different ratings. Overall, our results support the use of either faculty or student ratings of medical school lecturers, with the proviso that some quantitative differences do exist when lectures are viewed live versus on videotape.

The use of videotape review facilitates the acquisition of larger numbers of peer observational ratings. This addresses the problem of idiosyncratic rater bias that other investigators have noted in their discussions of faculty peer evaluations.[7,10,14] Videotape review provides a flexible method of observational evaluation of teaching for busy faculty who may be unable to attend live lectures.

Our findings, based in a medical school setting, mirror the majority of previous studies based in college and university settings that compared faculty and student ratings.[6,9,15] The considerable effort involved in obtaining faculty peer observational evaluations of teaching implies that peer evaluation will provide additional and potentially different information not obtained in the student evaluations. Yet on the three dimensions of teaching that we measured, there was no consistent difference in ratings between student and faculty peer evaluations. Of course, such an absence does not imply there is no utility to faculty evaluation of teaching. Faculty and students differ significantly in their expertise, perspectives, and content knowledge of the subject being taught.[2] Faculty peers may, for example, be better able to evaluate a lecturer's use of state-of-the-art subject content, a lecturer's selection of material that optimally promotes the overall course objectives, or a lecturer's pedagogical development from year to year. These structural characteristics of teaching may also be consistent across viewing conditions, and so easily amenable to video peer review. Our study suggests that to avoid duplication of efforts, faculty peer observation is best focused on different dimensions of teaching than those evaluated by students.

The results of our study are limited, for several reasons. First, sample size limits the power of our findings and may have obscured more meaningful relationships. We sampled only two faculty lectures, in order to use lecturers with equal exposure to the students in the course. At this time it would be difficult to recruit a cohort of medical students to view many lectures on videotape, and we believe this could detract from their educational experience. Third, the effect of viewing condition was not fully explored, since the faculty members did not view the live lectures. This may have prevented us from detecting important differences in how faculty rate lecturers. Finally, the theoretical basis of our measure was derived from studies involving non-medical school university courses rather than medical school courses, which may obscure important findings.

Overall, our results suggest that faculty peer evaluations of lectures are valid measures of teaching effectiveness but may be overvalued and yield little new information when compared with medical student evaluations using the same rating instruments. We believe peer evaluation should utilize faculty expertise to assess different dimensions of teaching than those evaluated by students. Videotaping lectures that faculty peers can later observe and evaluate in conjunction with course materials appears to be an effective and valid method of engaging larger numbers of busy clinical faculty in the peer evaluation of teaching effectiveness. Further studies may be useful in identifying those dimensions of medical students teaching for which faculty peer evaluation may add useful and unique information when compared with student evaluations.

Correspondence: Dr. Martin Leamon, University of California–Davis, Department of Psychiatry, 2230 Stockton Blvd, Sacramento, CA 95817; e-mail: (mhleamon@ucdavis.edu).

References

1. Nelson MS. Peer evaluation of teaching: an approach whose time has come. Acad Med. 1998;73:4–5.
2. Wilkerson L, Irby DM. Strategies for improving teaching practices: a comprehensive approach to faculty development. Acad Med. 1998;73:387–96.
3. McKeachie WJ. Student ratings: the validity of use. Am Psychologist. 1997;52:1218–25.
4. Irby DM. Peer review of teaching in medicine. J Med Educ. 1983;58:457–61.
5. Ballard M, Rearden J, Nelson L. Student and peer rating of faculty. Teaching of Psychology. 1976;3:88–90.
6. Cashin WE. Student Ratings of Teaching: The Research Revisited. IDEA Paper. Manhattan, KS: Center for Faculty Evaluation and Development, 1995.
7. Gromisch DS, Bamford JC Jr, Rous SN, Sall S, Rubin S. A comparison of student and departmental chairman evaluations of teaching performance. J Med Educ. 1972;47:281–4.
8. Stillman PL, Gillers MA, Heins M, Nicholson G, Sabers DL. Effect of immediate student evaluations on a multi-instructor course. J Med Educ. 1983;58:172–8.
9. Marsh HW, Roche LA. Making students' evaluations of teaching effectiveness effective: the critical issues of validity, bias, and utility. Am Psychologist. 1997;52:1187–97.
10. Feldman KA. Instructional Effectiveness of College Teachers as Judged by Teachers Themselves, Current and Former Students, Colleagues, Administrators, and External (Neutral) Observers. Research in Higher Education. 1989;30:137–94.
11. Heppner PP, Johnston JA. Peer consultation: faculty and students working together to improve teaching. Special feature: faculty development. Journal of Counseling & Development. 1994;72:492–9.
12. Hutchings P. Making Teaching Community Property: A Menu for Peer Collaboration and Peer Review. Washington, DC: American Association for Higher Education, 1996.
13. Jernstedt GC. Course Evaluation Modules. Hanover, NH: Center for Educational Outcomes, 1997.
14. Kulik JA, McKeachie WJ. The evaluation of teachers in higher education. In: Kerlinger FN (ed). Review of Research in Education. Itascs, IL: F. E. Peacock, 1975;3:210–40.
15. Greenwald AG. Validity concerns and usefulness of student ratings of instruction. Am Psychologist. 1997;52:1182–6.