Are online student evaluations of faculty influenced by the timing of evaluations?

John A. McNulty,1,2 Gregory Gruener,2,3,5 Arcot Chandrasekhar,1 Baltazar Espiritu,4 Amy Hoyt,3 and David Ensminger6

1Department of Cell and Molecular Physiology, 2Ralph P. Leischner Institute for Medical Education, and 3Office of Educational Affairs, Stritch School of Medicine, Loyola University, Maywood; Departments of 4Medicine and 5Neurology, Loyola University Health System, Chicago; and 6School of Education, Loyola University, Chicago, Illinois

Submitted 9 July 2010; accepted in final form 13 September 2010

STUDENT EVALUATIONS OF FACULTY are used for faculty promotion and tenure decisions as well as faculty development. Because of the importance of these events, student evaluations have become controversial with regard to their validity as a measure of teaching effectiveness (5, 7, 14, 16). Concerns have also been raised about the adequacy of methods used to evaluate teaching competencies in medical schools (2, 9). Some studies have suggested that student evaluations are not always related to the level of learning and include factors such as grade expectation, workload (6), and appearance of the instructor (16). Centra (3) also reported that courses rated as “too elementary” received lower evaluations. While students may not be able to judge all aspects of faculty performance, timeliness of the content, or the instructor’s knowledge of content, they can provide valuable feedback on instructional techniques and course design (4, 10). Similar results were found in medical student evaluations of faculty (10, 15).

The timing of student evaluations of preclinical courses also influenced the quality ratings of the faculty but only negligibly when the elapsed time between evaluations was within a 4-wk period (12). Because most student evaluations are completed at the end of a course, faculty who lecture only at the beginning of a 19-wk course do not have the opportunity to receive timely feedback. Moreover, delayed evaluations by students may not be as accurate. Accordingly, to provide more timely feedback to faculty and improve instructor evaluations by medical students, we gave students an opportunity to evaluate faculty during their physiology course concurrent with their lecture instead of waiting until the end of the course. The ability to evaluate faculty concurrently would remove barriers associated with end-of-course evaluations, such as forgetting specifics about individual lectures and lack of motivation to complete the evaluations. We then compared faculty evaluations from students who submitted them concurrently with those from students who waited to submit them at the end of a basic science course that covered a longer period of time (19 wk).

We extended the observations of McOwen et al. (12) to include analyses of student comments and examine possible associations of timing of evaluations with performance in the course. We hypothesized that concurrent evaluations would provide better discrimination among faculty and better commentary from the students.

METHODS

This study was a concurrent mixed model design with quantitative data being the dominant form of collection and analysis. Data were collected from first-year (M1) students during their integrated course in physiology. The class was 51% male and 49% female and ranged in age from 21 to 39 yr with an average age of 24 yr. Multiple faculty members (n = 16) taught blocks of varying content length (2–25 lectures) lasting 19 wk during the spring semester. Those faculty members whose lecture blocks were at the beginning of the course would be most affected by the lack of timely and accurate feedback when reviewed at the end of the course.

Evaluation instrument. A secure web-based application was developed to collect student evaluations of faculty. Students were given access to faculty evaluations beginning the first day of the course. They rated the 16 faculty members on a Likert scale of 1–5 (where 1 = strongly disagree and 5 = strongly agree) on the following three questions: 1) “The lecturer related the content to the learning objectives,” 2) “The lecturer communicated effectively,” and 3) “The
Students were encouraged to evaluate the faculty during their lecture blocks, although they had 2 wk after the course ended to complete their evaluations. Information on the process for submitting their evaluations was provided during orientation to the course and reminders were sent periodically during the course by e-mail. Students who submitted evaluations concurrently could edit their evaluations until finalized at the end of the course. As part of their professional obligations, students were required to complete evaluations of all faculty before finalizing, but they were allowed to enter “not applicable” if they did not attend the lecture(s) or did not have specific comments. All evaluations were anonymous to the faculty. The policies and requirements for submitting course and faculty evaluations were conveyed to the students during orientation. Those students who elected not to submit evaluations were invited to meet with the Dean of Educational Affairs to explain their noncompliance.

Data collection and analysis. Student evaluations of faculty were considered “concurrent” if they were posted to the server anytime between the first day of class and the last lecture given in the course. “End-of-course” evaluators were those who waited to submit their evaluations during the 2-wk period after the end of the course. Data were collected from the server logs, and individual student anonymity was maintained before further analyses. Student evaluations of the same 16 faculty members were collected from the previous year for year-to-year comparisons. The online evaluation items and rating scales were identical to those used in the study.

Qualitative data (i.e., students’ written comments) were evaluated through a rating process by four individuals (coauthors). This process involved scoring how “substantive” a student’s feedback was based on a scale of 1–4 that was weighted according to the amount of specific feedback in each comment. For instance, a score of 1 was given to comments that presented only one piece of feedback (e.g., “Great instructor”); a score of 4 was assigned to comments that included more than three pieces of feedback addressing different topics (e.g., “Clear and precise, good notes, good organization, easy to figure out the important material.”). All identifying information was removed from the comments before being scored. To determine interrater reliability, 100 of 539 rated comments were randomly selected. Reliability was calculated using Pearson’s correlation through paired correlations of scores for each rater against all other raters for these 100 ratings. Interrater reliability scores for how substantive the comments were ranged from $r(98) = 0.88–0.89$ for the four raters with a mean rating of $r = 0.89$. Student comments were also quantitated by counting the number of characters in each comment.

The degree of linear associations between two variables were determined using Pearson’s $r$. Student’s $t$-test was used for comparing differences between the means. $\alpha = 0.05$ was considered statistically significant. All analyses were conducted using Prism 4.0 software.

The final grades of students were used to measure performance in the course. These grades were calculated from five midterm exami-
nations. Once the data were entered into spreadsheets, the names of students were removed to preserve anonymity and confidentiality. The data collection process presented no harm to any of the participants. This study was reviewed by our Institutional Review Board and was exempted from further review and monitoring.

RESULTS

Are there differences in the degree to which students evaluated faculty concurrently versus at the end of the course? Over the study period, 38 of 144 students (26%) elected to submit evaluations concurrent with faculty lectures, 94 of 144 students (65%) waited until the end of the course, and 12 of 144 students (9%) elected not to submit any evaluations. Students began submitting their evaluations during the first week of the course and in every week thereafter. Almost half of the concurrent evaluations (150 of 305 evaluations) were submitted by week 13. Based on the data shown in Table 1, 38 students provided 305 concurrent evaluations, accounting for $\sim 15\%$ of the total evaluations submitted. The remaining evaluations from these 38 students were submitted at the end of the course. Most of the students (99%, 131 of 132 students) who submitted evaluations (concurrent and at the end of the course) evaluated all 16 lecturers. The majority of students (82%, 31 of 38 students) submitting evaluations concurrently elected to include comments compared with 62% (58 of 94 students) who submitted comments with their evaluations at the end of the course. Of 539 total evaluations with comments, only 15 (3%) evaluations were changed from the time the students submitted their evaluations on specific faculty and the time they finalized their evaluations at the end of the course.

The degree to which students discriminated between faculty was measured by the ranges of faculty scores (i.e., individual student ratings of the 16 faculty could range from 1 to 5). The majority of students (66%, 87 of 132 students) used either the entire Likert scale (1–5) or a nearly full scale (2–5) when evaluating all 16 faculty members. The full scale of 1–5 was used more frequently by concurrent evaluators (42%, 16 of 38 students) versus end-of-course evaluators (26%, 25 of 94 students). Only 8% (11 of 132 students) of the total evaluators used the same score (4 or 5) to rate all faculty. These 11 students submitted their evaluations at the end of the course.

Are there differences in faculty evaluations between students who evaluated concurrently versus after completion of the course? The average correlation of numerical scores for the three questions that students were asked to evaluate was $r(2,071) = 0.84$ ($P < 0.01$). This indicates that the students tended to rate faculty performance similarly across the three constructs assessed, (i.e., content knowledge, communication effectiveness, and learning facilitation) and suggests that there is a relationship between a faculty member’s performance on these three constructs. Accordingly, the scores for each faculty were averaged to produce a single score.

Table 1. Numbers of students and evaluations used in reporting the results

<table>
<thead>
<tr>
<th></th>
<th>Concurrent Evaluations</th>
<th>End-of-Course Evaluations</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of students submitting evaluations</td>
<td>38</td>
<td>94</td>
<td>132</td>
</tr>
<tr>
<td>No. of students submitting comments</td>
<td>31</td>
<td>58</td>
<td>89</td>
</tr>
<tr>
<td>No. of students who edited evaluations</td>
<td>7</td>
<td>Not applicable</td>
<td>7</td>
</tr>
<tr>
<td>Total no. of evaluations</td>
<td>305</td>
<td>1,768</td>
<td>2,073</td>
</tr>
<tr>
<td>Total no. of evaluations with comments</td>
<td>106</td>
<td>427</td>
<td>539</td>
</tr>
<tr>
<td>Percentage of evaluations with comments</td>
<td>35</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>Average final grade (SD)</td>
<td>86.7 (7.7)</td>
<td>83.9 (6.5)</td>
<td></td>
</tr>
</tbody>
</table>
Each of the 16 faculty members received concurrent evaluations, but the number of student evaluations for these faculty ranged from between 5 and 32 students (mean = 19 students, SD = 9.9). There was a significantly high correlation \([r(14) = 0.91, P < 0.01]\) between the average numerical scores faculty received from students who evaluated concurrently versus those who evaluated at the end of the course. The strength of this correlation indicated that student perception of faculty performance on the three constructs remained consistent regardless of when students completed the evaluation. This consistency of evaluation was also seen when year-to-year comparisons were made as faculty evaluations were highly correlated \([r(14) = 0.88, P < 0.01]\) with their numerical scores in the preceding year.

Do the frequency and quality of student comments on their evaluations vary between concurrent versus end-of-course evaluations? Students who submitted their evaluations concurrently tended to include comments more frequently (82%, 31 of 38 students) compared with those who evaluated at the end of the course (62%, 58 of 94 students; Table 1). The average number of faculty who received comments from early evaluators (mean = 5.6 faculty comments, SD = 4.4) was significantly higher \([t(141) = 3.08, P < 0.01, d = 0.44]\) compared with late evaluators (mean = 3.1 faculty comments, SD = 3.3). Furthermore, the comments of early evaluators were significantly longer \([t(102) = 2.9, P < 0.01, d = 0.44]\), comprising an average of 200 characters (SD = 131.1) compared with an average of 132 characters (SD = 89.6) for late evaluators.

Based on interrater reliability, an average “substantiveness score” was calculated for each individual by averaging the four rater scores for each participant. When the average rating scores for each group (i.e., early evaluators and end-of-course evaluators) were compared, comments from early evaluators were significantly more substantive \([t(537) = 3.2, P < 0.01, d = 0.36]\) with an average score of 2.22 (SD = 0.98) compared with an average score of 1.93 (SD = 0.96) for comments submitted at the end of course. The analyses indicated that those students completing concurrent evaluations provided more qualitative information about the lecturers, and the quality, in terms of the substantiveness of the comments, was significantly greater compared with students who completed their evaluations at the end of the course. This suggests that concurrent evaluations are more likely to provide faculty with more feedback about their instruction than end-of-course evaluations.

Are there associations between student evaluations of faculty and students’ final grade in the class? There were significant associations between how well students performed in the course and their evaluation tendencies. First, students who elected to evaluate concurrently had an average final grade that was significantly higher \([r(130) = 2.24, P < 0.03, d = 0.45]\) than those who evaluated at the end of the course (Table 1). Students who did not submit evaluations had a further reduction in their average grade (78.8, SD = 6.8) that was significantly lower than those students who submitted their evaluations at the end of the course \([r(104) = 2.68, P < 0.01, d = 0.85]\). Second, the average final grade of students who submitted one or more comments (mean = 85.8, SD = 5.7) was significantly higher \([r(130) = 2.73, P < 0.01, d = 0.47]\) compared with students who did not include comments with their evaluations (mean = 82.8, SD = 6.9). There were no associations between a student’s final grade and the average numerical evaluations they gave faculty \([r(130) = -0.03, P > 0.05]\).

**DISCUSSION**

The benefits of online student evaluations (12) were extended in this study to measure the effect of time as a variable. Although a minority of students (26%, 38 of 144 students) took advantage of the opportunity to evaluate faculty at or near the time of their lectures, it was noteworthy that the average faculty scores from these students were statistically similar to the evaluations of students who waited until the end of the course to fill in their evaluations. These results are consistent with those reported by McOwen et al. (12), who found that the timing of student evaluations had only a negligible impact on the scores given by students. We also found that once entered, students tended not to change their scores and comments regardless of the time they submitted their evaluations. Hence, first impressions appear to be important.

The significantly high correlations of faculty evaluations from year to year was further evidence of the stability and reliability of students’ critical assessments of teaching, as demonstrated by Krantz-Grod et al. (10) when student evaluations were compared over a similar 2-yr period. While no direct data were collected regarding faculty’s use of student evaluation information, the consistency of faculty ratings from year to year further suggests that faculty are not making significant adjustments to student concerns.

It is generally recognized that students’ comments have value to curriculum and faculty development (1, 13, 15). If student evaluations are to be used in a “dynamic” fashion by faculty to modify subsequent lectures, it is noteworthy that the comments of the early evaluators were significantly greater in number and length and that these comments were more substantive in that they included significantly more substantive comments. Concurrency evaluations could play a more beneficial role, even if only a minority of students participated (11). If the purpose of instructor evaluation is formative, these results suggest that having medical students complete evaluation forms immediately after viewing an instructor provides the best opportunity to gather not only the most qualitative information but also the richest qualitative information to inform the instructor’s teaching practice. There is evidence to support programmatic decisions that make use of fewer concurrent student evaluations of faculty lectures, with as few as 13 raters providing adequate \(\phi\) and G reliability but recognizing the rater as the primary source of total variance (11).

A significant finding of our study was the direct correlation between how well students performed in the course and their tendencies to evaluate early with more comments that were longer and more substantive. Possible reasons for this include greater motivation, engagement in the course, and valuing of the evaluation processes. The further reduction in average grade of those students who elected not to submit evaluations is consistent with this notion of lack of engagement in the course. It is important to note that the differences were found only in the willingness to provide qualitative comments for faculty and not the numerical rating, which suggests that those who receive higher grades are more willing to give the extra effort to provide qualitative feedback to instructors, whereas...
lower-scoring students seemed satisfied to just provide a quantitative rating. This result could be interpreted as those scoring higher are more vested in influencing the level of education they receive. This observation has important considerations in assessing the value of student evaluations when they are required to submit them, as is the case in many schools like ours. The assumption is that required evaluations may not be as insightful. The evidence from this study raises the question whether faculty scores would change significantly if student evaluations were voluntary. A relationship between grade expectations and student evaluations has been demonstrated (3, 8), but there was no association in our study between a student’s grade and the numerical evaluations they gave faculty.

The collection of data from only a single course at one institution may be a limitation of this study. More importantly, we did not determine to what degree students were noting their evaluations of faculty at the time of their lectures but did not submit this information until the end of the course. We also did not determine the degree to which faculty viewed their evaluations (early or late). The finding that evaluations of faculty did not change from year to year suggests that faculty who received poorer evaluations were not able to successfully address concerns raised by the students in earlier evaluations. Strategies are currently being considered to promote more effective use of student evaluations for enhancing curricular and faculty development.

ACKNOWLEDGMENTS

The authors thank Sabah Al-Janabi for assistance with collection of the data.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

REFERENCES