Pharmacy Education, December 2006; 6(4): 237–243

Relationships among student evaluations, instructor effectiveness, and academic performance†

STEPHEN D. PHIPPS[1], ROBERT S. KIDD[1], & DAVID A. LATIF[2]

[1]*Department of Biopharmaceutical Sciences, Shenandoah University Bernard J. Dunn School of Pharmacy, 1775 North Sector Court, Winchester, VA, 22601, USA, and* [2]*Department of Pharmaceutical Sciences, University of Charleston School of Pharmacy, 2300 MacCorkle Avenue, S.E., Charleston, WV 25304, USA*

(Received 15 December 2005; revised 19 June 2006; accepted 27 June 2006)

Abstract

This study was conducted to evaluate the relationships among students' grade expectations, students' actual grades, and students' evaluations of instructors. A total of 5399 individual student evaluations from 138 course offerings that were taught over four successive academic years were compiled and analyzed. The evaluation instrument included questions pertaining to course- and instructor-related items, as well as a question inquiring about the grade the student expected to receive in the course. Students' grades (expected and actual) were significantly correlated with the mean instructor evaluation score ($p < 0.01$ for both correlations). Also, there was a strong positive correlation ($r = 0.916$) between the mean course evaluation score and the mean instructor evaluation score ($p < 0.01$). Based on the results in this study, students' expected and actual course grades appear to be an influential factor in how they evaluate instructors. Additionally, the ability of students to discriminate between course evaluations and instructor evaluations is suspect.

Keywords: *Instructor effectiveness, student evaluations, student grades, course evaluation*

Introduction

Traditionally, student evaluations have been the primary mechanism employed to assess both course and instructor effectiveness (Barnett & Matthews, 1998). While several schools of pharmacy use other assessment methods (e.g. peer, expert, administrative, and self-evaluation), student evaluations are used by 100% of US schools and colleges of pharmacy (Barnett & Matthews, 1998). From an administrative standpoint, student evaluations are commonly used to directly assess faculty members, which include rewarding excellence in teaching, course load allocation, faculty performance reviews, promotion and tenure, and merit raises. Because of the summative nature of student evaluations at many schools of pharmacy, coupled with questions about whether students have a sufficient knowledge base in either course content or pedagogical theory to provide valid evaluations, faculty may harbor legitimate concerns about the weight of student evaluations, the priority accorded them, and the resultant impact on decisions that affect the faculty member.

The reliability and validity of student evaluations has been the subject of much debate. Some data support the reliability and validity of using student evaluations to evaluate teacher effectiveness (McKeachie, 2002) while other data suggest student evaluations can be biased (Canaday, Mendelson, & Hardin, 1978; Carline & Scher, 1981; West, 1988). Criticism of student evaluations includes sampling bias (self-selection of those completing a course evaluation; Carline & Scher, 1981) and related response rates, timing bias (when the evaluation is administered during the course; Canaday et al., 1978), and temporal stability (West, 1988). Perhaps the most

Correspondence: S. D. Phipps, Department of Biopharmaceutical Sciences, Shenandoah University Bernard J. Dunn School of Pharmacy, 1775 North Sector Court, Winchester, VA, 22601, USA. Tel.: + 011 540 545-7313. Fax: + 011 540 665-1283. E-mail: sphipp2@su.edu

†The present research was conducted within the Department of Biopharmaceutical Sciences at Shenandoah University Bernard J. Dunn School of Pharmacy.

ISSN 1560-2214 print/ISSN 1477-2701 online © 2006 Informa UK Ltd.
DOI: 10.1080/15602210600886175

significant factor regarding pharmacy faculty evaluations are data suggesting a practically significant relationship in the positive direction between grade expectations and course evaluations (Kidd & Latif, 2004). In that study, the investigators analyzed a total of 5399 individual student evaluations from 138 courses taught over several years at one school of pharmacy (Kidd & Latif, 2004). The authors concluded that the relationship between students' grade expectations for a course and how they evaluated the course was highly correlated ($p < 0.001$).

Since course evaluations assess global aspects of the course (course resources, exams/assessment, complexity, methods of evaluations, etc.), they may not accurately reflect instructor-related items such as presentation clarity, depth, organization, thoroughness, and stimulating student interest, motivation and learning. Even if the criticisms were adequately addressed and faculty concerns regarding reliability and validity were alleviated, another factor to consider is that course evaluations often include both course- and instructor-specific items. As such, course evaluations may not be reflective of instructor evaluations. Given the multifaceted utilization and impact of student evaluations on faculty issues, it would seem appropriate to either isolate or separate these domains (course- and instructor-specific items) or to account for these distinct factors, thus improving the specificity and accuracy of the evaluation instrument when evaluating teacher effectiveness. Incorporating both elements (course- and instructor-specific items) may erroneously and nonspecifically equate course effectiveness and teacher effectiveness

Another issue pertaining to the reliability of student evaluations is the relationship between students' academic achievement in the course and the corresponding evaluation form that they complete. As discussed above, in a previous study we evaluated the relationship between actual and expected course grades and course evaluations and found that there was a significant positive correlation between course grades and course evaluation scores (Kidd & Latif, 2004). One limitation identified in that study was that course assessment, not teacher assessment, was used for the correlation analyses (i.e. course evaluations may not be generalized to instructor effectiveness). The current study was designed to evaluate the hypotheses that teacher evaluation scores are also strongly positively correlated with student grade expectations and actual grades. To increase the precision of this investigation, several hypotheses are advanced to evaluate four sub-domains of instructor effectiveness (lecture content, presentation/style, learning, and student contact) and the relationship among these instructor qualities and actual grade, overall course evaluation, and overall teacher evaluation.

Materials and methods

Prior to the study, the research protocol was approved by the Human Subjects Review Board of Shenandoah University. This investigation used a convenience sample and was a blinded retrospective record review of course and instructors evaluations. A power analysis, based on an estimated moderate effect size, used a 0.80 convention to determine the proper sample size. Based on this analysis, using an *a priori* 0.05 significance level, 80 evaluated courses would be needed for this investigation (Hair, Anderson, Tatham, & Black, 1998). Student evaluations from 138 required and elective courses over four academic years were compiled and analyzed. The course evaluations were comprised of 5399 individual student evaluations from first-, second-, and third-year pharmacy courses taught between the fall semester of 1999 and the spring semester of 2003. These four academic years were selected for three reasons. The first reason was to obtain a sufficient number of courses as determined by the power analysis. A second reason for examining this time period is that the same evaluation instrument was used during this period (prior to this period, a different instrument was used). Finally, this time period and these data were selected for inclusion in this study because they coincide with data analyzed in a previous study in which course-specific parameters were examined in the context of actual and expected grades (Kidd & Latif, 2004). The present investigation builds on the previous one by examining instructor-specific items within those same courses.

The total number of students completing each course and the corresponding mean grade (on a 4.0 scale) for each course was obtained from the university registrar. Students completed the course evaluations near the conclusion of the course, most often within the last week of the course and prior to the final course examination. The mean course evaluation scores were calculated as the numerical average of the nine questions in the course evaluation form (Appendix 1) that related to the students' perceptions of the course. Responses to each of those questions was in the form of a 5-letter Likert scale anchored at A, strongly agree and E, strongly disagree. The numerical means for those questions were calculated by assigning a number value to each response as follows: $A = 5$, $B = 4$, $C = 3$, $D = 2$ and $E = 1$.

The mean instructor evaluation scores were calculated in a similar manner to the mean course evaluation scores, except that the 13 questions that pertained to the students' assessment of instructor related items were examined rather than the nine items targeting the course evaluation (Appendix 1). On the evaluation instrument, instructor characteristics are categorized into four sub-domains of teacher effectiveness. Domain mean scores were calculated using only the three or four questions in the respective domain. The mean grade that students expected to receive was calculated as the numerical mean of the course grade question (Appendix 1, question 5) using a conventional scale of $A = 4.0$, $B = 3.0$, $C = 2.0$, $D = 1.0$, and E (fail) $= 0$. Pearson r correlation analysis was used to examine the various hypothesized relationships between instructor evaluations, expected grades, actual grades, course evaluations, and specific instructor effectiveness domains.

For those courses in which multiple instructors were involved, the instructor evaluation score was determined by calculating the mean score of each of those instructors evaluated in the course.

Finally, a third set of correlations was performed to analyze the relationships between the individual instructor domains and the corresponding course grade, and also to evaluate the relationship between those instructor domains and the course evaluations. SPSS v. 11 was used to evaluate the data for statistical significance.

Results

During the four academic years included in this study, 217 didactic courses were offered at the school of pharmacy (Fall 1999 through Spring 2003). Of those, course evaluations were conducted in 138 course offerings (64%). Since individual courses were offered and evaluated multiple times over this time period, the 138 course offerings were comprised of 58 individual courses. Of the 7474 students who completed these 138 course offerings, 5399 (72%) completed a course evaluation. The completed course evaluations were the basis of the study analyses and these sample data are summarized in Table I. The mean instructor evaluation score was 4.22. This overall mean score is comprised of the four domains of instructor effectiveness. Table II provides a summary of the instructor effectiveness domain scores.

The first set of correlation analyses was conducted using a Pearson r correlation test to determine the significance of the following relationships: (1) the correlation between instructor evaluations and expected grade, (2) the correlation between instructor evaluations and actual grade, and (3) the correlation between instructor evaluations and course evaluations.

The correlation between the students' expected grade and the instructor evaluation scores revealed the following: the mean expected grade was 3.26 ± 0.56; the mean instructor evaluation score was 4.22 ± 0.52 and the resulting Pearson correlation coefficient was $r = 0.46$. As depicted in Figure 1, the Pearson r correlation indicated that students' grade expectations were significantly correlated to instructor evaluation scores ($p < 0.01$).

The correlation between students' actual grades and the instructor evaluation scores resulted in a Pearson correlation coefficient of $r = 0.434$ (mean = 3.15 ± 0.64). As shown in Figure 2, the Pearson r correlation indicated that students' actual grades were significantly correlated to instructor evaluation scores ($p < 0.01$).

The relationship between course evaluations and instructor evaluations resulted in a mean course evaluation score of 4.03 ± 0.53, a mean instructor evaluation score of 4.22 ± 0.52, and a Pearson correlation coefficient of $r = 0.916$. Figure 3 shows that course evaluations and instructor evaluations are significantly correlated ($p < 0.01$). A second set of correlations was conducted to evaluate the relationship between instructor effectiveness and instructor evaluations scores. Specifically, four correlation tests

Table I. Summary of sample data.

Variable	
Number of years assessed	Four academic years (1999–2003)
Number of course offerings evaluated	138
Number of student evaluations	5399
Mean course evaluation score	4.03*
Mean instructor evaluation score	4.22*
Mean expected grade	3.26†
Mean actual grade	3.15†
Professional years assessed	P-1–P-3

*Likert scale anchored at 5, strongly agree and 1, strongly disagree. †$A = 4$, $B = 3$, $C = 2$, $D = 1$, E (fail) $= 0$.

Table II. Summary of instructor characteristics.

Instructor domain*	Mean score†
Lecture content	4.28
Presentation/style	4.26
Learning	4.17
Student contact	4.24

*See Appendix I for individual evaluation questions comprising the respective domains. †Likert scale anchored at 5, strongly agree and 1, strongly disagree.

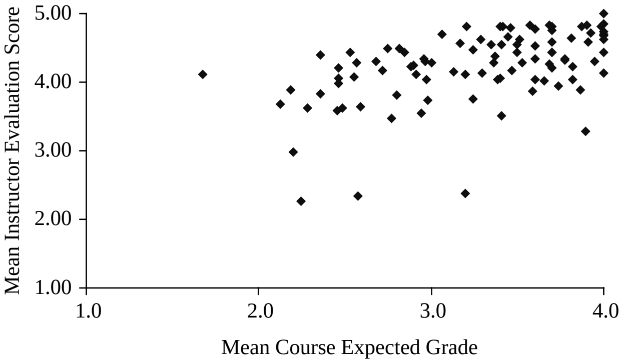

Figure 1. Mean course expected grade versus mean instructor evaluation score for 138 individual course offerings ($r = 0.460$).

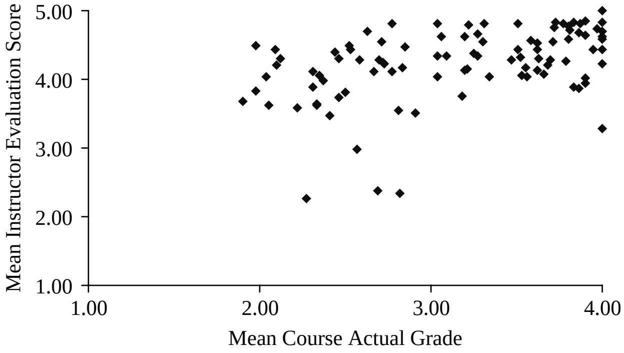

Figure 2. Mean course actual grade versus mean instructor evaluation score for 138 individual course offerings ($r = 0.434$).

were performed to assess the relative rank and degree of significance of the four domains of instructor effectiveness (i.e. lecture content, presentation style, learning, and student contact) and their individual relationship to instructor evaluation scores. Each individual instructor domain was significantly correlated to the instructor evaluation scores. Table III summarizes the results of the four correlation analyses between instructor scores and effectiveness domains.

Finally, a third set of correlations was performed to analyze the relationships between the individual instructor domains and the corresponding course grade, and also to evaluate the relationship between those instructor domains and the course evaluations. The results of the correlation analyses indicated that each one of the four instructor domains was significantly correlated to both the actual course grade ($p < 0.01$) and to the course evaluation ($p < 0.01$). These relationships are summarized in Tables IVa and IVb.

Discussion

The first two analyses examined the relationship between course grade (expected and actual) and instructor evaluations. In both analyses, the correlations were both statistically significant. These results indicate that students' perceptions and evaluations of faculty are significantly influenced by their grade. This raises several concerns, including calling into question the objectivity of those completing the evaluation forms. Additionally, if grades provide a significant impact on instructors' evaluations, then the validity of what the instrument is intended to assess may be compromised (i.e. is the evaluation instrument truly assessing instructor characteristics, or is it assessing instructor characteristics skewed by an incorporated achievement bias?). The present data would seemingly support the latter. Despite reports regarding the validity of student ratings, Fish (2004) has suggested that the assumptions underlying student evaluations is more reflective of "customer satisfaction than with the soundness of one's pedagogy".

Given the significance accorded to student evaluations with regard to their subsequent utilization and impact on faculty (e.g. promotion, tenure and merit raises), the data suggest that better instructor evaluations might be attained if higher grades are assigned. The appeal of more favorable evaluations might provide a biased incentive to lower expectations or grading standards. Grade inflation has been reported, not as an isolated event, but as a continuing trend. Granberry and Stiegler (2003) recently reported that pharmacy graduates' grade point averages have increased by about 1% per year during a 20-year period, but did not see increases in preprofessional grade point average or Pharmacy College Admission Test scores in the same time period. If student evaluations were not as heavily weighted or as intimately linked to faculty performance assessments, one could speculate that the prevalence and extent of grade inflation might be significantly lower.

Course evaluations and instructor evaluations were highly correlated. A possible explanation for this high correlation is that students are unable to separate their satisfaction with the course and the instructor. If students liked the instructor, they may also rate the course favorably regardless of how effectively it was conducted. Similarly, student dissatisfaction with the course (e.g. policies, cost of textbook and schedule) might be projected onto the instructor and influence how they score the instructor. When this highly correlated course-instructor relationship is considered in conjunction with the apparent impact of student grades, it appears firstly, that students are not categorically discerning between course and instructor evaluations, and secondly, that academic achievement (expected and actual) influences student perceptions of both the course and instructor.

Other correlation analyses indicated that each one of the instructor effectiveness domains (lecture content, presentation/style, learning, and student contact) was significantly correlated with the overall instructor

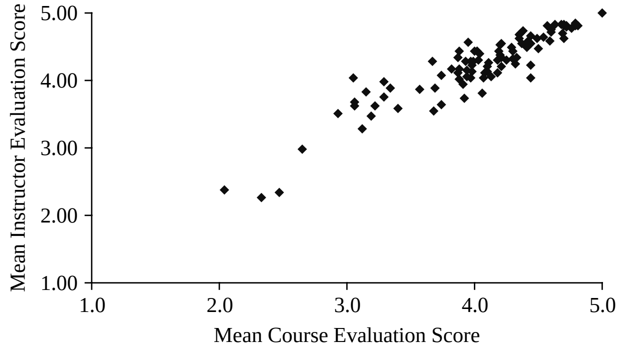

Figure 3. Mean course evaluation score versus mean instructor evaluation score for 138 individual course offerings ($r = 0.916$).

Table III. Summary of instructor domains and instructor scores.

Instructor domain* (independent variable)	Instructor score (dependent variable)	*p*-value of correlation
Lecture content mean score†	Mean instructor evaluation score†	$p < 0.01$ ($r = 0.956$)
Presentation/style mean score†	Mean instructor evaluation score†	$p < 0.01$ ($r = 0.960$)
Learning mean score†	Mean instructor evaluation score†	$p < 0.01$ ($r = 0.975$)
Student contact mean score†	Mean instructor evaluation score†	$p < 0.01$ ($r = 0.924$)

*See Appendix I for individual evaluation questions comprising the respective domains. †Likert scale anchored at 5, strongly agree and 1, strongly disagree.

evaluation score (Table III), with the actual course grade, and with the course evaluation. Intuitively, it was not unexpected to discover that a significant relationship exists between an instructor's effectiveness and the associated overall instructor evaluation. It seems logical that students would tend to favorably evaluate those instructors who, among other evaluated traits, are motivating, helpful, explanatory, and accessible. The present data appear to support that conclusion. Another significant correlation that was identified was between the effectiveness domains and the overall course grade (Table IVa). As one would expect, good pedagogy beneficially impacts learning and retention of course content. Not as apparent, however, are the reasons underlying the significant relationship between instructor effectiveness domains and the course evaluation. One possible reason for this may be similar to that mentioned above, which is that students may transpose or equate instructor- and course-specific items. To better dissociate these parameters will be a challenging task, and may involve increasing the number and/or specificity of items that are queried on the evaluation tool.

Other evaluation methods are available that may provide more reliable and consistent results. Peer review, while less commonly employed, may circumvent the biases that are associated with student evaluations (Speer & Elnicki, 1999). It can be argued that assessing competence and excellence in teaching might be better evaluated by a scholarly and professional peer (Fish, 2004), while simultaneously alleviating the instructor's incentive to lower standards and inflate grades (Wilson, 1998a). As an intellectual and scholarly endeavor, "teaching, like research, should be peer reviewed" (Wilson, 1998b). However, a similar question arises with regard to peer review, which is "Can reviewers critically evaluate their colleagues?" (Speer & Elnicki, 1999). A less frequent evaluation mechanism is self-appraisal or self-rating. Barnett & Matthews (1998) reported that only twelve schools of pharmacy utilize this method of teaching evaluations. More recently, these investigators reported the reliability of faculty self-evaluation and recommended that this method be incorporated as part of the instructor evaluation process (Barnett, Matthews, & Jackson, 2003).

While the sample size (number of courses and evaluations) in this study was relatively large, one limitation of the present investigation is that data from only one school of pharmacy were analyzed. Another limitation is that the evaluation instrument was developed internally and has not been validated.

Table IVa. Summary of instructor domains and course grade.

Instructor domain* (independent variable)	Course grade (dependent variable)	*p*-value of correlation
Lecture content mean score†	Mean course grade‡	$p < 0.01$ ($r = 0.323$)
Presentation/style mean score†	Mean course grade‡	$p < 0.01$ ($r = 0.363$)
Learning mean score†	Mean course grade‡	$p < 0.01$ ($r = 0.434$)
Student contact mean score†	Mean course grade‡	$p < 0.01$ ($r = 0.396$)

†Likert scale anchored at 5, strongly agree and 1, strongly disagree. ‡*A*, 4; *B*, 3; *C*, 2; *D*, 1; *E* (fail), 0. *See Appendix 1 for individual evaluation questions comprising the respective domains.

Table IVb. Summary of instructor domains and course evaluation scores.

Instructor domain* (independent variable)	Course evaluation score (dependent variable)	*p*-value of correlation
Lecture content mean score†	Mean course evaluation score†	$p < 0.01$ ($r = 0.912$)
Presentation/style mean score†	Mean course evaluation score†	$p < 0.01$ ($r = 0.871$)
Learning mean score†	Mean course evaluation score†	$p < 0.01$ ($r = 0.943$)
Student contact mean score†	Mean course evaluation score†	$p < 0.01$ ($r = 0.880$)

†Likert scale anchored at 5, strongly agree and 1, strongly disagree. *See Appendix 1 for individual evaluation questions comprising the respective domains.

However, there was consistent utilization of the student evaluation instrument (Appendix 1) over the course of the four years that were analyzed.

Conclusions

This study found a strong positive correlation between students' grade expectations and actual grades and instructor evaluation scores. It also demonstrated a strong positive correlation between course evaluations and instructor evaluations. Finally, every one of the four evaluated instructor domains (i.e. lecture content, presentation style, learning and student contact) was positively correlated to the actual course grade. Since student course and instructor evaluations are commonly used to evaluate faculty in schools and colleges of pharmacy, it is imperative that we understand the factors that influences these evaluations. In addition, it is essential for schools and colleges of pharmacy to evaluate the proper and improper use of students' course and instructor evaluations. Based upon the findings in this study, sole dependency on student evaluations to assess teacher effectiveness is tenuous at best. Rather than to abandon student evaluations altogether, perhaps they should be utilized as one component of a more comprehensive instructor evaluation process, including peer evaluation, content-expert evaluation, and self-evaluation. To confirm these results and answer the questions raised, these findings should be evaluated at other schools and colleges of pharmacy.

References

Barnett, C. W., & Matthews, H. W. (1998). Current procedures used to evaluate teaching in schools of pharmacy. *American Journal of Pharmaceutical Education*, *62*, 388–391.

Barnett, C. W., Matthews, H. W., & Jackson, R. A. (2003). A comparison between student ratings and faculty self-ratings of instructional effectiveness. *American Journal of Pharmaceutical Education*, *67*(4), 117.

Canaday, S. D., Mendelson, M. A., & Hardin, J. H. (1978). The effect of timing on the validity of student ratings. *Journal of Medical Education*, *53*, 958–964.

Carline, J. D., & Scher, M. (1981). Comparison of course evaluations by random and volunteer student samples. *Journal of Medical Education*, *56*, 122–127.

Fish, S. (2004). Who's in charge here? The evaluation of teaching by students amounts to a whole lot of machinery with a small and dubious yield. *Chronicle of Higher Education*, *51*(22), C2.

Granberry, M. C., & Stiegler, K. A. (2003). Documentation and analysis of increased grade point averages at a college of pharmacy over 20 years. *American Journal of Pharmaceutical Education*, *67*, 1–5.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis*, 5th ed. (pp. 11–13). Upper Saddle River, NJ: Prentice Hall.

Kidd, R. S., & Latif, D. A. (2004). Student evaluations: Are they valid measures of course effectiveness? *American Journal of Pharmaceutical Education*, *68*, Article 61.

McKeachie, W. J. (2002). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers*, 11th ed. (pp. 326–327). Boston, MA: Houghton Mifflin Company.

Speer, A. J., & Elnicki, D. M. (1999). Assessing the quality of teaching. *American Journal of Medicine*, *106*(4), 381–384.

West, R. F. (1988). The short-term stability of student ratings of instruction in medical school. *Medical Education*, *22*, 104–112.

Wilson, R. (1998a). New research casts doubt on value of student evaluations of professors. *Chronicle of Higher Education*, *44*(19), A12.

Wilson, R. (1998b). Project seeks to help colleges use peer review to evaluate teaching. *Chronicle of Higher Education*, *44*(19), A14.

Appendix 1. Course evaluation instrument

Bernard J. Dunn School of Pharmacy course evaluation

Please take a few minutes to seriously consider & complete this form. Your responses will be used as a part of the process of faculty evaluation of this professor(s) and this course.

A, strongly agree, B, agree, C, neutral, D, disagree, E, strongly disagree

Course evaluation

1. The resources (e.g. textbook, notes and slides) used in this course contributed to my learning.
2. Integrated teaching was effectively used in this course (if applicable). If the question is not applicable to this course, please do not select a choice.
3. I understood the subject matter of this course.
4. The content of the laboratory of recitation was a worthwhile part of this course (if applicable). If the question is not applicable to this course, please do not select a choice.

Student expectations of the course

5. Grade I expect to receive in this course. If you expect to receive the grade of F in this course, please select choice E.

Examinations/grades

6. Exams/Assignments accurately assessed what was taught in this course.
7. Complexity and length of course assignments were reasonable.
8. Methods of evaluation were fair.
9. Feedback on evaluations was valuable.
10. Graded assignments and examinations were returned in a timely fashion.

Instructor

Lecture content

11. Lecture content adequately addressed objectives.
12. The instructor used current information in his/her lectures for this course.
13. Lecture topics were discussed in sufficient depth.

Presentation/style

14. The instructor had an organized style of presentation.
15. Teaching methods used by the instructor were appropriate for the material presented.
16. The instructor explained difficult material clearly.
17. The instructor spoke audibly and clearly.

Learning

18. The instructor motivated me to do my best work.
19. Students were encouraged to contribute to class learning.
20. The instructor stimulated interest in the material.

Student contact

21. The instructor was accessible when students had problems.
22. The instructor was helpful when students had problems.
23. The instructor encouraged independent learning.