Journal of Business Administration Online
HTTP://JBAO.ATU.EDU

Fall 2002, Vol. 1 No. 2

STUDENT EVALUATION OF INSTRUCTORS: A MEASURE OF TEACHING EFFECTIVENESS OR OF SOMETHING ELSE?

Kevin H. Mason, Arkansas Tech University
Robert R. Edwards, Arkansas Tech University
David W. Roach, Arkansas Tech University

Introduction

Student evaluations of teaching performance have been used since the 1920s, yet their validity, the techniques used to administer them, and the purposes for which they are used remain controversial (Marsh, 1984,1987; Spencer & Flyr, 1992; Wachtel, 1998). In the early years, individual instructors usually made the decision whether or not to use student evaluations, designed their own evaluation instruments, and were the only ones who saw the results. During the 1970s, however, many universities began requiring student evaluations, standardizing evaluation instruments, and scoring the evaluation results for performance appraisal purposes (Centra, 1993). Student feedback is now a component in the formal faculty performance appraisal systems of most universities (Adams, 1997; Lersch & Greek, 2001; Trout, 1997).

Faculty opinions on the use of student evaluations range widely, but a considerable proportion of researchers conclude that the evaluations provide a valid and reliable method for judging teaching effectiveness (Centra, 1977, 1993; Cohen, 1980, 1981; Koon & Murray, 1995; Marsh & Dunkin, 1992; Ramsden, 1991; Seldin, 1984, 1993). Further, Marsh (1987) asserts that student evaluations are the only indicator of teaching effectiveness whose validity has been rigorously and thoroughly established.

Despite the research that supports the validity of student evaluations, many individuals express reservations about their use in faculty performance appraisal systems (Adams, 1997; Chandler, 1978; Dowell & Neal, 1982; Goldman, 1993; Tata, 1999; Zoller, 1992). A common concern is the possibility that factors other than teaching effectiveness influence the evaluation scores. These include the procedures used to administer the evaluations (Seldin, 1993), the anonymity of the evaluators (Blunt, 1991; Feldman, 1979), whether the course is required or elective (Brandenburg et al., 1977; Feldman, 1978; Scherr & Scierr, 1990), the class meeting time (Centra, 1993; Koushka & Kuhn, 1982), whether or not the course requires quantitative reasoning (Cashin, 1990, 1992; Feldman, 1978; Ramsden, 1991), the course workload (Ryan et al., 1980), the personal characteristics of the instructor (Marsh & Dunkin, 1992; Radmacher & Martin, 2001), and the students' prior interest in the course subject area (Marsh & Cooper, 1981; Prave & Baril, 1993).

There is evidence that a positive correlation exists between a student's anticipated course grade and the student's overall evaluation of the instructor. Koshland, (1991), Lersch & Greek, (2001), Nimmer & Stone, (1991), and Vasta & Sarmiento, (1979) conclude that instructors with lenient grading standards receive higher overall ratings, and Chacko (1983) suggests that strict grading standards also lead students to rank the instructor lower on evaluation components (such as self-reliance and attitude toward students) that are unrelated to judgments about grading fairness.

Relatively little research has been conducted on student reactions to the evaluation process and their potential effect on the assigned ratings. Abbott et al. (1990) observe that students often complain about the frequency with which they are asked to complete evaluation forms, a phenomenon noted earlier by Brandenburg et al. (1979), who also questioned whether students take the evaluations seriously. Marlin (1987) states that students tend to view the evaluations as a chance to "let off steam" and Jacobs (1987) reports that 40 percent of her student respondents said they were aware of other students plotting to "get back at" an instructor by collectively assigning low ratings. Trout (2000) provides anecdotal evidence of students rewarding easy-grading instructors with high evaluation scores and notes the devastating effect that even a few disengaged students can have on quantitative evaluation scores simply by giving an instructor the lowest possible scores on all evaluated dimensions.

In addition to validity and reliability issues, some researchers question the importance assigned to student evaluations in faculty performance appraisal systems. Although Rice et al., (2000) accept student evaluations as a valuable tool for assessing and improving classroom teaching, they argue that these evaluations do not capture information about long-term instructor and course effectiveness. With Adams (1997) and Ruben (1997), they conclude that any assessment of teaching effectiveness should rely on multiple perspectives collected from various university stakeholders.

The results of several studies provide a general consensus about some apparent dimensions of teaching effectiveness (Braskamp et al., 1981; Feldman, 1997; Marsh & Dunkin, 1997; Murray, 1997; Perry, 1997; Solomon et al., 1964; Wotruba & Wright, 1975). These include the teacher's (1) knowledge of the subject matter, (2) preparation and organization of the course, (3) sensitivity to and concern for students, (4) fairness in grading, (5) helpfulness, (7) elocutionary skills, and (8) class management, as well as (9) the effectiveness of instructional aids (textbook, etc.), and (10) the clarity of course objectives.

Purpose of The Study

The purpose of this exploratory study was to identify factors that may influence university students' evaluations of instructor effectiveness.

Methods

Students enrolled in 37 sections of 12 different Management and Marketing courses at a moderately sized southwestern university evaluated their instructors near the end of semester-long courses. The evaluations were prepared anonymously, without the presence of the

instructors, and the students were informed that the instructors would not be permitted to see the results until grades for the respective classes were officially reported. A total of 797 student evaluations were collected on four instructors over a three-and-a-half-year period

The university's standard student evaluation questionnaire allows students to evaluate 10 common dimensions of instructor performance chosen from the literature on the subject. Table 1 displays the specific questionnaire items included on the questionnaire. Additionally, the form asks students to report their age (17-24, 25-40, over 40), their approximate overall grade (A, B, C, D), the grade they anticipate in the particular course (A, B, C, D), and whether the course is required for their major. Space is also provided for unscored written comments to be made, at the evaluating student's option.

Table 1: Items Rated By Students To Evaluate Class Instructors

	Items Rated*
1.	The instructor was knowledgeable in this field.
2.	The instructor effectively presented the content of the course.
3.	The instructor was well prepared for each class.
4.	The instructor was available to provide assistance outside of class.
5.	The instructor evaluated my work in this course fairly.
6.	The instructor's overall performance as a teacher was excellent.
7.	The class time was valuable in helping my understanding.
8.	If a textbook was required in the course, it was useful.
9.	The other instructional aids, if used, were beneficial.
10.	The instructor is fluent in English.

* Students provided their level of agreement with each item on a 5-point scale where higher values represented higher levels of agreement.

Dependent Variable

Using the individual students' level of agreement/disagreement with each statement, we computed an instructor's overall composite evaluation score for each section by averaging the students' assigned scores across all ten items, resulting in a maximum possible score of 50 for the most favorable rating. This composite served as the dependent variable for our study.

Independent Variables

For each course section, we assessed each student's overall grade average, anticipated grade for the course, and age group. In addition, we recorded whether the student provided written comments (and, if so, whether the comments were generally favorable or unfavorable). Additionally, we noted the faculty member's academic rank and tenure status, whether the

course was required, the time of day the course was scheduled (morning or afternoon), the length of each class meeting (50 or 80 minutes), and the level of the course (Freshman, Sophomore, Junior, or Senior). These factors served as independent variables.

Results

Correlation analysis was conducted to examine the relationships between the items rated by students on the faculty evaluation questionnaire. A correlation matrix (Table 2) demonstrates the existence of multicollinearity among the ratings of these measures. It would appear from Table 2 that the ratings are not very discerning with regard to assessing different dimensions of performance. Rather, it appears that faculty were judged on overall performance and scored similarly on all dimensions. Hence, the overall composite score appears to be a reasonable representation of the overall faculty evaluation ratings.

Table 2: Correlation Matrix For Dependent Measures (n = 797)
Correlations Significant at .0001 denoted by *

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Composite
Item 1	1.0										
Item 2	.59*	1.0									
Item 3	.65*	.64*	1.0								
Item 4	.50*	.53*	.50*	1.0							
Item 5	.53*	.68*	.57*	.61*	1.0						
Item 6	.58*	.78*	.60*	.64*	.74*	1.0					
Item 7	.52*	.70*	.54*	.53*	.60*	.73*	1.0				
Item 8	.31*	.38*	.33*	.30*	.40*	.36*	.35*	1.0			
Item 9	.43*	.51*	.46*	.52*	.50*	.55*	.60*	.43*	1.0		
Item 10	.44*	.34*	.41*	.36*	.36*	.37*	.28*	.24*	.25*	1.0	
Composite	.71*	.84*	.74*	.75*	.82*	.88*	.82*	.57*	.74*	.49*	1.0

T- tests and Analysis of variance (ANOVA), coupled with Tukey HSD pair-wise tests where appropriate, were conducted to determine which, if any, of the independent variables were related to the composite faculty evaluation rating. Several of the independent variables were not found to have a significant relationship to the dependent variable. These included the student's age, the students' overall grade average, the course level (consistent with Cooper et al., 1982), instructor rank (also consistent with Cooper et al., 1982), instructor tenure, and whether the course was required. However, significant effects were observed for the remaining independent variables (expected grade, time of day, length of class meeting, the provision of written comments, and the tone of such comments -- favorable or unfavorable).

The most powerfully related factor to faculty evaluations was the student's expected grade in the course ($F = 25.2$, d.f. = 3, 771, $p < .0001$). Tukey HSD tests for differences in faculty evaluations revealed that those students who expected to get an "A" in the course assigned significantly higher faculty evaluations than those who expected a lower grade. In the same manner, those students who expected to get a "B" in the course assigned significantly higher faculty evaluations than those expecting lower grades. Interestingly, while students who

expected to receive a “C” in the course assigned higher ratings than those who expected to get a “D”, these differences were not significant. These means are displayed in Table 3.

Table 3: Mean Composite Ratings By Students’ Expected Course Grade

Expected Grade	Mean Instructor Rating*
A	46.24
B	44.82
C	41.92
D	39.35

* Maximum total evaluation = 50

Course offering time was significantly related to faculty ratings, as well. Table 4 shows the mean ratings sorted by the time of day the course was offered. As shown in Table 4, the T-test indicates that instructors of afternoon courses were rated higher than instructors of morning classes.

Table 4: T-Test for Differences in Mean Ratings Across Class Offering Times

Mean Instructor Rating For Morning Classes*	Mean Instructor Rating For Afternoon Classes*	T-Value For Difference	P-Value
44.16 (n = 551)	45.34 (n = 243)	-.271	.0068

* Maximum total evaluation = 50

Significant effects were also observed for the length of the class period. Specifically, based upon pairwise T-tests, faculty evaluations collected from course sections that met in 50-minute intervals were significantly higher than evaluations collected from sections that met in 80-minute intervals (see Table 5).

ANOVA results also indicated a significant relationship between the faculty evaluations and the written comments factor. That is, the faculty composite ratings differed significantly across students, depending upon their response to the written comment section of the evaluation. ($F = 81.82$, d.f. = 2, 790, $p < .0001$). Specifically, based upon the Tukey HSD results, the mean composite rating provided by those who wrote favorable comments (47.33) was significantly higher than the evaluations provided by those who wrote unfavorable comments (35.83) or who did not write any comments at all (44.11). Furthermore, those who did not write comments at all provided more favorable faculty evaluations than those students who wrote unfavorable comments. Table 6 provides these means.

Table 5: T-Test for Differences in Mean Ratings Across Lengths of Class Meetings

Mean Instructor Rating For 50-Minute Classes*	Mean Instructor Rating For 80-Minute Classes*	T-Value For Difference	P-Value
45.02 (n = 624)	42.72 (n = 170)	4.75	<.0001

* Maximum total evaluation = 50

Table 6: Mean Composite Ratings By Written Response Factor

Written Comments	Mean Composite Faculty Evaluation *
Favorable	47.33
None	44.11
Unfavorable	35.83

* Maximum total evaluation = 50

Note: Based upon the Tukey HSD test, all means differ significantly.

Discussion

These results indicate that student evaluations of teacher performance are strongly and directly related to student grade expectations. This outcome is consistent with the conclusions of Koshland (1991), Lersch & Greek (2001), Nimmer & Stone (1991), and Vasta and Sarmiento (1979). They also support Chacko's (1983) suggestion that grading leniency is positively related to student scoring of other effectiveness dimensions. This supports the conclusion of these researchers that student evaluations may be more a measure of their satisfaction with the rigor of grading than a valid measure of teaching effectiveness

The assignment of significantly different ratings in morning and afternoon suggest two obvious, but conflicting possibilities. The first is that the instructors rated highest by the students simply taught a larger proportion of their classes in the afternoon than the instructors who received lower evaluation scores. The second possible explanation is that the students had a preference for afternoon classes that was strong enough to influence their ratings significantly. Based on our experience in academic advising, we suspect that the second possibility is unlikely – when a course has both morning and afternoon meeting times, the morning sections (except 8 a.m.) tend to reach capacity levels sooner and/or have larger enrollments than the afternoon sections.

A similar dilemma is presented by the significant difference between the evaluation scores in the 50-minute and 80-minute class lengths. As above, the apparent explanation is either the more frequent scheduling of higher-rated instructors to 50-minute classes or a student preference for that class length that influences their evaluations of instructors. As with the afternoon versus morning discussion, we consider the former explanation to be more plausible.

The significant relationships between evaluation scores and the existence/nature of optional written comments are less surprising. One might reasonably expect that students who express high levels of satisfaction or dissatisfaction through their evaluations have stronger feelings about the instructor's performance than students who assign ratings in the middle of the scale. A larger proportion of those with strong feelings can reasonably be expected to take the extra effort to expend the effort necessary to provide written comments.

Conclusion

The high levels of multicollinearity among the scores of the ten rating dimensions suggest that students are unable to discern varying levels of effectiveness among the dimensions. Instead, they appear to form an overall conclusion about the instructor's performance and assign similar scores on all dimensions. If this is true, one, (or a few) of the dimensions (or some unrated element) is so important to the students that it overwhelms objective assessment of other dimensions.

While the results of our study suggest a clear relationship between students' anticipated course grades and their overall evaluations of instructors, additional research is needed. Using individual students as the unit of examination, this exploratory study provided a sufficient population (797) for statistical analysis. All the measurements came from the evaluations of only four instructors, however, and one might logically assert that the data should really be analyzed for individual instructors. This, and the examination of other independent variables that might affect student evaluation scores (e.g., pre-existing student perceptions about a particular instructor or course), is a subject for future studies. Research on this subject is further complicated by the need to maintain student anonymity. We acknowledge also that our use of an unweighted composite evaluation score may not be sophisticated enough for conclusive results.

Given the widespread use of student evaluations to assess faculty teaching effectiveness, the validity of the evaluations may need continuing confirmation. It is possible that the validity established for this type of measurement in past decades may not apply as well in today's campus environment. In addition, faculty performance appraisal needs to include input from other sources, including alumni, peers, and administrators. We concur with Adams (1997),Rice et al. (2000) and Ruben (1997) that any legitimate assessment of teaching effectiveness must incorporate multiple perspectives provided by various university stakeholders.

Student evaluations may be one useful source of information about a teacher's performance, but our results, and those of some earlier studies, suggest that the evaluations are influenced by factors other than teaching effectiveness. Unless the effects of these factors can be isolated, it appears that faculty performance appraisal should not rely excessively on evaluations provided anonymously by students.

References

Abbott, R.D., Wulf, D.H., Myquist, J.D., Ropp, V.A. & Hess, C.W. (1990). Satisfaction with processes of collecting student opinions about instruction: The student perspective. *Journal of Educational Psychology*, 82, 201-206.

Adams, J.V. (1997). Student evaluations: The ratings game. *Inquiry*, 1(2), 6-10.

Blunt, A. (1991). The effects of anonymity and manipulated grades on student ratings of instructors. *Community College Review*, 18, 48-54.

Brandenburg, D.C., Slinde, J.A., & Batista, E.E. (1977). Student ratings of instruction: Validity and normative interpretations. *Research in Higher Education*, 7, 67-78.

Brandenburg, D.C., Braskamp, L.A., & Ory, J.C. (1979). Considerations for an evaluation program of instructional quality. *CEDR Quarterly*, 12, 8-12.

Braskamp, L.A., Ory, J.C., & Pieper, D.M. (1981). Student written comments: Dimensions of instructional quality. *Journal of Educational Psychology*, 73, 65-70.

Cashin, W.E. (1990). Students do rate different academic fields differently. In M. Theall & J. Franklin (Eds.), *Student Ratings of Instruction: Issues for Improving Practice, New Directions for Teaching and Learning, No. 43*. San Francisco: Jossey-Bass.

Cashin, W.E. (1992). Student ratings: The need for comparative data. *Instructional Evaluation and Faculty Development*, 12, 146.

Centra, J.A. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal*, 14, 17-24.

Centra, J.A. (1993). *Reflective faculty evaluation*. San Francisco: Jossey-Bass.

Chacko, T.I. (1983). Student ratings of instruction: A function of grading standards. *Educational Research Quarterly*, 8, 19-25.

Chandler, J.A. (1978). The questionable status of student evaluations of teaching. *Teaching of Psychology*, 5,150-152.

Cohen, P.A. (1980). Using student ratings feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education*, 13, 321-341.

Cohen, P.A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, 51, 281-309.

Cooper, P.J., Stewart, L.P., & Gudykunst, W.B. (1982). Relationship with instructor and other variables influencing student evaluations of instruction. *Communication Education*, 30, 308-315.

Dowell, D.A. & Neal, J.A. (1982). A selective review of the validity of student ratings of teaching. *Journal of Higher Education*, 53, 51-62.

Feldman, K.A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education*, 9, 199-242.

Feldman, K.A. (1979). The significance of circumstances for college students= ratings of their teachers and courses. *Research in Higher Education*, 10, 149-172.

Feldman, K.A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R.P. Perry & J.C. Smart (Eds.), *Effective Teaching in Higher Education: Research and Practice* (pp. 368-395). New York: Agathon Press.

Goldman, L. (1993). On the erosion of education and the eroding foundations of teacher education (or why we should not take student evaluation of faculty seriously. *Teacher Education Quarterly*, 20, 57-64.

Jacobs, L.C. (1987). *University Faculty and Students= Opinions of Student Ratings*. Indiana Studies in Higher Education, #55. Bloomington, IN: Bureau of Evaluation and Testing, Indiana University.

Koon, J. & Murray, H.G. (1995). Using multiple outcomes to validate student ratings of overall teacher effectiveness. *Journal of Higher Education*, 66, 61-81.

Koshland, D.E. (1991). Teaching and Research. *Science*, 251, 249.

Koushki, P.A. & Kuhn H.A.J. (1982). How reliable are student evaluations of teaching? *Engineering Education*, 72, 362-367.

Lersch, K.M. & Greek, C. (2001). Exploring the beliefs surrounding student evaluations of instruction in criminology and criminal justice undergraduate courses. *Journal of Criminal Justice Education*, 12(2), 283-299.

Marlin, J.W., Jr. (1987). Student perception of end-of-course evaluations. *Journal of Higher Education*, 58, 704-716.

Marsh, H.W. & Cooper, T.L. (1981). Prior subject interest, students' evaluations, and instructor effectiveness. *Multivariate Behavioral Research*, 16, 82-104.

Marsh, H.W. (1984). Students' evaluation of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76, 707-754.

Marsh, H.W. (1987). Students' evaluation of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research,* 11, 253-288.

Marsh, H.W. & Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J.C. Smart (Ed.), *Higher Education: Handbook of Theory and Research, Vol. 8,* (pp. 143-233). New York: Agathon Press.

Marsh, H.W. & Dunkin, M.J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R.P. Perry & J.C. Smart (Eds.), *Effective Teaching in Higher Education: Research and Practice,* (pp. 241-320). New York: Agathon Press.

Murray, H.G. (1997). Effective behaviors in the college classroom. In R.P. Perry & J.C. Smart (Eds.), *Effective Teaching in Higher Education: Research and Practice* (pp. 171-204). New York: Agathon Press.

Nimmer, J.G. & Stone, E.F. (1991). Effects of grading practices and time of rating on student ratings of faculty performance and student learning. *Research in Higher Education,* 32, 195-215.

Perry, R.P. (1997). Teaching effectively: Which students? What methods? In R.P. Perry & J.C. Smart (Eds.). *Effective*
Teaching in Higher Education: Research and Practice (pp. 154-168). New York: Agathon Press.

Prave, R.S. & Baril, G.L. (1993). Instructor Ratings: Controlling for bias from initial student interest. *Journal of Education for Business,* 68, 362-366.

Radmacher, S.A. & Martin, D.J. (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *Journal of Psychology,* 135, 259-268.

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The course experience questionnaire. *Studies in Higher Education,* 16, 129-150.

Rice, R.E., Stewart, L.P., & Hujber (2000). Extending the domain of instructional effectiveness assessment in student evaluations of communication courses. *Communication Education,* 40, 253-266.

Ruben, B. (1997). *Excellence in Higher Education: A Guide to Self-Assessment, Strategic Planning and Improvement*. Dubuque, IA: Kendall-Hunt.

Ryan, J.J., Anderson, J.A., & Birchler, A.B. (1980). Student evaluations: The faculty responds. *Research in Higher Education,* 12, 317-333.

Scherr, F.C. & Scierr, S.S. (1990). Bias in student evaluations of teacher effectiveness. *Journal of Education for Business,* 65, 356-358.

Seldin, P. (1984). *Changing Practices in Faculty Development.* San Francisco: Jossey-Bass.

Seldin, P. (1993). The use and abuse of student ratings of professors. *Chronicle of Higher Education*, 39(46), p. A40.

Solomon, D., Rosenberg, L., & Bezdek, W. (1964). Dimensions of teacher behavior. *Journal of Experimental Education*, 55, 23-30.

Spencer, P.A. & Flyr, M.L. (1992). The formal evaluation as an impetus to classroom change: Myth or reality? (Research/Technical Report, Riverside, CA).

Tata, J. (1999). Grade distributions, grading procedures, and students' evaluation of instructors: A justice perspective. *Journal of Psychology*, 133 (3), 263-271.

Trout, P. (1997, Sept./Oct.). What the numbers mean: Providing a context for numerical student evaluations of courses. *Change: The Magazine of Higher Learning*, 25-30.

Trout, P. (2000, Apr. 21). Teacher evaluations. *Commonweal*, 127(8), 10-11.

Vasta, R. & Sarmiento, R.F. (1979). Liberal grading improves evaluations but not performance. *Journal of Educational Psychology*, 71, 207-211.

Wachtel, H.K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education*, 23, 191-211.

Wotruba, T.R. & Wright, P.L. (1975). How to develop a teacher-rating instrument: A research approach. *Journal of Higher Education*, 46, 653-663.

Zoller, U. (1992). Faculty teaching performance evaluation in higher science education: Issues and implications (a cross-cultural case study). *Science Education*, 76, 673-684.