AN ANALYSIS OF STUDENT EVALUATIONS OF COLLEGE TEACHERS

Milam Aiken, University of Mississippi, University, MS

Del Hawley, University of Mississippi, University, MS

Mahesh Vanjani, Texas Southern University, Houston, Texas

ABSTRACT

An analysis of 4,325 undergraduate Business students' evaluations of their instructors' teaching performances shows that, as expected, course grades are positively correlated and class sizes are negatively correlated. However, other factors such as a teacher's preparation and speaking ability might be significantly correlated only through a grade-related halo effect. Further study is needed to identify additional contributors to measures of teacher assessment.

INTRODUCTION

The subject of students' ratings of teacher effectiveness is one of the most researched areas in education with over 2000 articles and books written on the topic over the past 70 years (Ory & Ryan, 2001), and yet, there is still no complete agreement on the effectiveness of this evaluation process. Some studies (e.g., Theall & Franklin, 2001)

have shown that there are consistently high correlations between students' ratings of the amount learned in the course and their overall ratings of the teacher and the course. However, serious doubts have been cast on the content of typical end-of-semester college teacher evaluation questionnaires, and many instructors have the view that students only know whether instructors are likeable, not whether they are knowledgeable, and students know whether lectures are entertaining, not whether the content is accurate and up to date (Cahn, 1987).

Here, we describe an analysis of one semester's student evaluations of teachers at a business school in the Southern region of the United States that seeks to determine which factors influence an instructor's overall performance rating. A correlation analysis shows that many of the questions on the survey are significantly related to performance, and a multi-linear regression model with these variables obtains an accuracy of about 94%. Finally, a new measure known as the Teacher Perception Index (TPI) is introduced that could be superior to the traditional evaluation of an instructor's ability.

BACKGROUND

Some research has shown that several factors have very little or no impact on evaluations of instructor performance, including personality traits (Erdle, Murray, & Rushton, 1985), instructor gender (Feldman, 1993), class size (Centra, 1993; Feldman, 1984), students' academic ability as measured by GPA (Theall & Franklin, 1990), and class time of day (Franklin, 2001). However, other factors have been found to be significantly related, including whether or not the course is an elective (Marsh, 1984), whether or not the students are majoring in the area (Feldman, 1978), the level of the course (Bausell & Bausall,

1979), and course difficulty (Cohen, 1981; Marsh & Roche, 2000; Te-Shang, 2000).

Perhaps most research has focused on the relation between students' expected grades and ratings of teacher performance. Some studies have found little or no relationship (e.g., Bacon & Novotny, 2002; Bilimoria 1995; Centra, 2003; Griffin, 2004; Webster 1990), while others have found a significantly positive relationship (e.g., Clayson & Haley 1990; Engdahl, Keating, & Perrachione, 1993; Heckert, Latier, Ringwald-Burton, & Drazen, 2006; Krautmann & Sander, 1999; Millea & Grimes, 2002; Nowell & Alston, 2007). In one meta-study of the phenomenon (Aleamoni, 1999), results showed that 24 studies found no significant relationship, while 37 studies found a significant, positive relationship with a median correlation of 0.14. However, some studies' correlations have ranged as high as 0.43 (Cohen, 1981), especially when students knew their final grades before rating their instructors (Abrami, Perry, & Leventhal, 1982).

Although several studies have sought to determine the effects of individual factors such as class size and instructor gender on teacher performance or have sought to find the correlations among several independent variables, only a few have tried to determine the effect of many factors together (e.g., DeCanio, 1986; Krautmann & Sander, 1999). For example, one study (Marks, 2000), used structural equation modeling (LISREL) to study the effect of difficulty, organization, fairness of grading, instructor liking, and perceived learning on teacher performance. Results in the study were mixed, calling into question the quality of the questionnaire used.

METHOD AND ANALYSIS

In an attempt to study the effect of variables used in our own end-of-semester teacher evaluation questionnaire, we retrieved evaluations from 111 sections of Management, Marketing, Finance, Management Information Systems, and Production and Operations Management courses taken during the fall semester of 2007 at the School of Business Administration in a medium-sized, Southern university. These summary evaluations represented a total of 4,325 individual evaluations made by undergraduate students. The evaluations were voluntary, and therefore, have some self-selection bias. However, a majority of students from each course section usually decided to complete the Web-based questionnaire (shown in Appendix I). If the students opted to complete the survey, they were allowed to register for the next year's classes earlier than usual, providing a little positive incentive. It should be noted, however, that students were not required to answer every question in the survey.

Braskamp and Ory (1994) identify six factors commonly found in student rating forms: course organization and planning, clarity and communication skills, teacher student interaction and rapport, course difficulty and workload, grading and examinations, and student self-learning. Although we do not believe the questionnaire's construct validity has ever been tested, it does contain examples of these six factors. However, this potential lack of construct validity is a limitation to the study.

Summary results with means and standard deviations are shown in Table 1.

Table 1: Evaluation Summary (N= 111)

Variable	Mean	Std Dev
N	38.964	28.009
q1	3.429	0.435
q2	3.312	0.613
q3	2.954	0.692
q4	3.338	0.421
q5	3.131	0.442
q6	3.163	0.569
q7	3.333	0.345
q8	3.241	0.449
q9	3.084	0.435
q10	2.346	0.573
q11	2.928	0.561
score	2.442	0.495
avg	3.115	0.370

N = number of students in the course section answering the survey
q1 – q10 (see Appendix I - scale 0 bad to 4 good)
score = average grade for the section (F = 0, … A = 4)
avg = average of q1 through q11

Using a difference of means t-test, results showed that all measures q1 – q11 were significantly above the neutral value of 2 at $\alpha = 0.05$, indicating that those students who responded were generally satisfied. Class scores were also significantly above 2 ("C"). Instructor preparation (Q1) and enthusiasm (Q4) received the highest marks, while course difficulty (Q10) received the lowest mark. The administration places a high value on course difficulty and rigor, but apparently, most students did not believe the classes were extremely hard.

There were several interesting, large, significant correlations among the variables (Table 2). For example, enthusiastic instructors (Q4) were also well prepared (Q1) (R=0.73) and spoke clearly (Q2) (R=0.59). Other variables could naturally be considered complementary and correlated, e.g. speaking clearly (Q2) and responding to students' questions in class (Q7) (R=0.75), preparedness (Q1) and appropriate examinations (Q5) (R=0.74), and returning assignments on time (Q6) and preparedness (Q1) (R=0.69). Also, as might be expected, there was a moderate positive correlation between the appropriateness of the examinations (Q5) and class score (R=0.43), and a negative correlation between course difficulty (Q10) and score (R = -0.44). Only two variables were not significantly correlated with teacher performance (Q11); class size and course difficulty (Q10). Of the significant variables related to this performance, only returning assignments on time (Q6), the quality of the book (Q9), and the class score had correlations less than or equal to 0.50. The class score had only a moderate correlation with teacher performance (R=0.35), somewhat above the median R of 0.14 reported in (Aleamoni, 1999). Finally, the average of variables Q1 to Q10 had a very high correlation with Q11 (teacher performance) (R = .95).

Table 2: Correlation Analysis (R/p-value)

	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	score	avg
N	-0.26	-0.27	-0.22	-0.21	-0.17	-0.21	-0.28	-0.06	-0.23	-0.11	-0.17	0.09	-0.27
	0.01	<.01	0.02	0.03	0.08	0.03	<.01	0.55	0.01	0.24	0.07	0.34	<.01
q1		0.67	0.86	0.73	0.74	0.69	0.83	0.72	0.35	0.08	0.86	0.20	0.92
		<.01	<.01	<.01	<.01	<.01	<.01	<.01	<.01	0.38	<.01	0.03	<.01
q2			0.76	0.59	0.67	0.32	0.75	0.54	0.35	-0.15	0.72	0.23	0.77
			<.01	<.01	<.01	<.01	<.01	<.01	<.01	0.11	<.01	0.01	<.01
q3				0.76	0.89	0.50	0.92	0.71	0.46	-0.15	0.94	0.37	0.93
				<.01	<.01	<.01	<.01	<.01	<.01	0.11	<.01	<.01	<.01
q4					0.71	0.40	0.82	0.65	0.36	-0.02	0.82	0.26	0.82
					<.01	<.01	<.01	<.01	<.01	0.86	<.01	0.01	<.01
q5						0.40	0.85	0.64	0.54	-0.19	0.89	0.43	0.86
						<.01	<.01	<.01	<.01	0.05	<.01	<.01	<.01
q6							0.42	0.65	0.12	0.24	0.50	0.19	0.65

							<.01	<.01	0.21	0.01	<.01	0.05	<.01
q7								0.67	0.51	-0.10	0.92	0.32	0.92
								<.01	<.01	0.29	<.01	<.01	<.01
q8									0.22	0.06	0.75	0.40	0.81
									0.02	0.54	<.01	<.01	<.01
q9										-0.26	0.47	0.35	0.48
										0.01	<.01	<.01	<.01
q10											-0.08	-0.44	0.07
											0.40	<.01	0.46
q11												0.35	0.95
												<.01	<.01
score													0.31
													<.01

N = number of students in the course section answering the survey
q1 – q10 (see Appendix I), score = average grade for the section (F = 0, … A = 4)
avg = average of q1 through q11

Next, the data were split into two subsets of 55 and 56 records each with a least square multi-linear regression analysis of one set as the training sample applied to the second set as the testing or holdout sample. Using Q11 (teacher performance) as the dependent variable, results showed a mean absolute percentage error (MAPE) of 7.4% for the first test set and 4.3% for the second test set (5.85% average). The SAS General Linear Model (GLM) model results for all data observations are shown in Table 3, and the multi-linear regression parameters are listed in Table 4.

In comparison, an analysis using naïve estimates (i.e., the estimate for the dependent variable in the testing sample is forecasted to be the same as the mean value of the dependent variable in the training sample), showed an MAPE of 41.5% for the first set and 24.2% for the second set (32.85% average).

Table 3: SAS GLM Model Results

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	32.483	2.707	122.35	< .001
Error	98	2.168	0.022		
Corrected Total	110	34.651			

R-Square	Coeff Var	Root MSE	Q11 Mean
0.937	5.081	0.149	2.928

Source	**DF**	**Type I SS**	**Mean Square**	**F**	**Pr > F**
score	1	4.282	4.282	193.550	<.0001
q10	1	0.241	0.241	10.890	0.001
q9	1	4.990	4.990	225.560	<.0001
q8	1	13.757	13.757	621.830	<.0001
q7	1	7.322	7.322	330.950	<.0001
q6	1	0.041	0.041	1.850	0.177

q5	1	0.983	0.983	44.440	<.0001
q4	1	0.149	0.149	6.730	0.011
q3	1	0.501	0.501	22.630	<.0001
q2	1	0.000	0.000	0.000	0.969
q1	1	0.132	0.132	5.940	0.017
N	1	0.085	0.085	3.850	0.053

Source	**DF**	**Type III SS**	**Mean Square**	**F**	**Pr > F**
N	1	0.085	0.085	3.850	0.053
q1	1	0.121	0.121	5.450	0.022
q10	1	0.009	0.009	0.410	0.522
q2	1	0.000	0.000	0.010	0.935

N = number of students in the course section answering the survey
q1 – q10 (see Appendix I - scale 0 bad to 4 good)
score = average grade for the section (F = 0, … A = 4)
avg = average of q1 through q10

Table 4: Regression model

(Q11 teacher performance - dependent variable)

Parameter	Estimate	Error	t value	Pr > |t|
Intercept	-1.295	0.272	-4.760	<.0001
N	0.001	0.001	1.960	0.053
q1	0.206	0.088	2.330	0.022
q2	0.003	0.038	0.080	0.935
q3	0.267	0.074	3.600	0.001
q4	0.168	0.062	2.690	0.008
q5	0.248	0.077	3.200	0.002
q6	-0.040	0.043	-0.930	0.353
q7	0.248	0.137	1.810	0.073
q8	0.156	0.058	2.670	0.009
q9	0.044	0.043	1.010	0.314
q10	0.021	0.032	0.640	0.522
score	-0.020	0.039	-0.510	0.610

N = number of students in the course section answering the survey
q1 – q10 (see Appendix I - scale 0 bad to 4 good)
score = average grade for the section (F = 0, … A = 4)
avg = average of q1 through q10

DISCUSSION

As in other schools during other semesters, students might have allowed their satisfaction with their expected grade to influence their ratings on other measures. For example, if a student realizes that she will receive a poor grade in the class, she might rate the instructor lower on speaking clearly or his enthusiasm, regardless of whether or not it was true, solely out of spite. On the other hand, a student who expects to receive a high grade in the class might give high ratings on everything, despite the facts.

We believe this grade-halo effect might well be the case for students in this study. There was a moderate (R=0.35) correlation between score and teacher performance, and a slightly lower (R=0.31) correlation between the score and the average of Q1 through Q10. Despite the very high correlation between the Q1 through Q10 average and instructor performance (Q11) (R=0.95) and the 94.15% performance forecasting accuracy of the multi-linear model, we believe that these independent variables were adjusted to fit the students' overall satisfaction with the class, and determined by their expected grade or some other factor, rather than the instructors' performances determined by the ratings of the independent variables. This grade-related halo effect has been used as an explanation in other studies for high marks on the legibility of the instructor's writing, the instructor's audibility, and the quality of classroom facilities (Greenwald & Gillmore, 1997).

As another illustration of the unreliability of the questionnaire and students' responses, we investigated a few questions in greater detail. For example, of all the questions in the survey, we believe Q6 (returning assignments and examinations in a reasonable period of time) is the least subjective, and anecdotal student responses illustrate the large variation and possible bias in the entire evaluation process.

In one course, 66% of the students who received an average score of 4.0 ("A") said materials were returned on time, and 33% said "not applicable." In comparison, 55% of those students receiving an average score of 3.62 said the materials were returned on time, 33% said "almost always", and 11% said "not applicable." However, the instructor always returned graded tests at the next class period, and assignments were always graded and returned via electronic mail within a few hours at most. Further, there is no reason why students should have marked "not applicable" if they were conscientiously responding to the survey. In the same class, there was no book (Q9), but only 44% reported the question as "not applicable." Similar results were found in other classes. Thus, the students might be concentrating on a few questions such as teacher performance (Q11) and spending relatively less time on others.

Course difficulty and rigor (Q10) is a factor the School considers as important in its curriculum. While moderately large (R = -0.44), it is somewhat surprising that the score / difficulty correlation is not even more negative. This could be due to any one or a combination of three factors:

1. The student is bragging. That is, the student might claim the course is difficult, but he also claims to be so smart that he received a good grade.

2. The student thinks grading is not fair. The student thinks he should have gotten a better grade because the material was not difficult, but due to unfair grading policies, this was not possible.

3. Inattention or lack of consideration for the question. Not much thought was given to answering the question correctly.

Because class scores influence evaluations of teacher performance, and the administration wants courses to be challenging, the School has implemented a Teacher Perception Index (TPI) that is a simple product of teacher performance multiplied by course difficulty (0 to 16 scale). The advantage of this performance measure is that it does not appear to be correlated with class score ($R = -0.118$, $p = 0.217$), but it is still significantly and positively correlated with many of the other variables in the survey such as instructor preparedness and enthusiasm.

CONCLUSION

This analysis of one semester's undergraduate student evaluations of teachers at a business school replicates some earlier studies' results (e.g., there was a moderate correlation between students' class scores and their teacher performance marks) and also provides some possible reasons why students answered the evaluation survey as they did. Class grades can influence perceptions, but other factors not included in the survey such as instructor likability, course content, and many other variables could also play a major role.

REFERENCES

Abrami, P., Perry, R., & Leventhal, L. (1982). The relationship between student personality characteristics, teacher ratings, and student achievement. *Journal of Educational Psychology*, 74, 111-125.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, 13, 153-166.

Bacon, D., & Novotny, J. (2002). Exploring achievement striving as a moderator of the grading leniency effect. *Journal of Marketing Education,* 24(1), 4-14.

Bausell, R., & Bausell, C. (1979). Student ratings and various instructional variables from a within-instructor perspective. *Research in Higher Education,* 11, 167-177.

Bilimoria, D. (1995). Modernism, postmodernism, and contemporary grading practices. *Journal of Management Education,* 19, 440-57.

Braskamp, L., & Ory, J. (1994). *Assessing Faculty Work: Enhancing Individual and Institutional Performance*. San Francisco, Jossey-Bass.

Cahn, S. (1987). Faculty members should be evaluated by their peers, not by their students. *Chronicle of Higher Education,* October 14, B2.

Centra, J. (1993). *Reflective Faculty Evaluation*. San Francisco, Jossey-Bass.

Centra, J. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Journal of Research in Higher Education,* 44(5), 495-518.

Clayson, D., & Haley, D. (1990). Student evaluations in Marketing: What is actually being measured? *Journal of Marketing Education,* 12, 9-17.

Cohen, P. (1981). Student ratings of instruction and student achievement: A meta-analysis of multi-section validity studies. *Review of Education Research,* 51(3), 281-309.

DeCanio, S. (1986). Student evaluations of teaching: A multinominal Logit approach. *The Journal of Economic Education,* 17(3), 165-176.

Engdahl, R., Keating, R., & Perrachione, J. (1993). Effects of grade feedback on student evaluation of instruction. *Journal of Management Education,* 17, 174-84.

Erdle, S., Murray, H., & Rushton, J. (1985). Personality, classroom, behavior, and college teaching effectiveness: A path analysis. *Journal of Educational Psychology,* 77, 394-407.

Feldman, K. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education,* 21(1) 45-116.

Feldman, K. (1993). College students' views of male and female college teachers: Part II – Evidence from students' evaluations of their classroom teachers. *Research in Higher Education,* 34, 151-211.

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

Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. *New Directions for Teaching and Learning,* 87, 85-100.

Greenwald, A., & Gillmore, G. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist,* 52, 1209-1217.

Griffin, B. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology,* 29(4), 410-425.

Heckert, T., Latier, A., Ringwald-Burton, A., & Drazen, C. (2006). Relations among student effort, perceived class difficulty appropriateness, and student evaluations of teaching: Is it possible to "buy" better evaluations through lenient grading? *College Student Journal,* 40(3), 588-596.

Krautmann, A., & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review,* 18(1), 59-63.

Marks, R. (2000). Determinants of student evaluations of global measures of instructor and course value. *Journal of Marketing Education,* 22(2), 108-119.

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

Marsh, H., & Roche, L. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology,* 92, 202-228.

Millea, M., & Grimes, P. (2002). Grade expectations and student evaluation of teaching. *College Student Journal,* 36 (4), 582-91.

Nowell, C., & Alston, R. (2007). I thought I got an A! Overconfidence across the Economics curriculum. *Journal of Economic Education,* 38(2), 131-143.

Ory, J., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In M. Theall, P. Abrami, and L. Mets (eds.), The Student Ratings Debate: Are they Valid? How Can We Best Use Them? *New Directions for Institutional Research,* 109, San Francisco, Jossey-Bass.

Te-Sheng, C. (2000). Student ratings: What are teacher college students telling us about them? *Annual Meeting of the American Educational Research Association,* New Orleans, LA, April 24-28.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places – A search for truth or a witch hunt in student ratings of instruction? In The Student Ratings Debate: Are they Valid? How Can We Best Use Them? Theall, P., Abrami, L. and Lisa, M. (Eds.) *New Directions in Educational Research,* 109, San Francisco, Jossey-Bass.

Theall, M., and Franklin, J. (1990). Student Ratings of Instruction: Issues for Improving Practice. *New Directions for Teaching and Learning,* 43. San Francisco, Jossey-Bass.

Webster, C. (1990. Evaluation of Marketing professors: A comparison of student, peer, and self-evaluations. *Journal of Marketing Education,* 12, 11-17.

APPENDIX I

TEACHER EVALUATION QUESTIONNAIRE

Overall Performance Weight:

A = 4 B = 3 C = 2 D = 1 E = 0

1. Was the instructor well organized and prepared for class sessions?

A. Always B. Almost Always C. Usually D. Sometimes E. Rarely

2. Did the instructor speak clearly and distinctly?

A. Always B. Almost Always C. Usually D. Sometimes E. Rarely

3. Did the instructor's classroom lectures and activities help you in learning the material?

A. Always B. Almost Always C. Usually D. Sometimes E. Rarely

4. Which best describes the instructor's attitude toward the subject matter?

A. Great B. C. Seems to D. E. Doesn't

enthusiasm	**Enthusiastic**	**like subject**	**Indifferent**	**like subject**

5. Were the examinations appropriate for assessing mastery of the course material?

A. Always	**B. Almost Always**	**C. Usually**	**D. Not often enough**	**E. Not applicable**

Overall Performance Weight:

A = 0 B = 2 C = 3 D = 4 E = Not Applicable

6. Did the instructor return assignments and examinations in a reasonable period of time?

A. No	**B. Usually**	**C. Almost Always**	**D. Always**	**E. Not applicable**

Overall Performance Weight:

A = 4 B = 3 C = 2 D = 1 E = 0

7. How helpful were the instructor's responses to students' questions in class?

A. Very helpful	**B. Helpful**	**C. Unhelpful**	**D. Rarely took questions**

Overall Performance Weight:

A = 0 B = 2 C = 3 D = 4 E = Not Applicable

8. If you needed assistance from the instructor outside of class, could you make satisfactory arrangements for a timely meeting?

A. No	**B. Usually**	**C. Almost always**	**D. Always**	**E. Not applicable**

Overall Performance Weight:

A = 4 B = 3 C = 2 D = 1 E = Not Applicable

9. How beneficial were the books required for this course?

A. Very beneficial	**B. Beneficial**	**C. Marginal**	**D. Not helpful**	**E. Not applicable**

Overall Performance Weight:

A = 4 B = 3 C = 2 D = 1 E = 0

10. How would you rate the difficulty level of this course, compared to other courses you have taken so far?

A. Extremely difficult	**B. Very difficult**	**C. Difficult**	**D. Average**	**E. Easy**

11. How would you rate the instructor's overall performance in this course?

A. Superior	B. Excellent	C. Good	D. Marginal	E. Poor

AUTHORS

Milam Aiken is a Professor and Chair of Management Information Systems in the School of Business Administration at the University of Mississippi. He has published over 100 articles in journals including Information & Management; ACM Transactions on Information Systems; IEEE Transactions on Man, Machines, and Cybernetics; International Journal of Knowledge Engineering and Software Engineering; and Decision Support Systems, and has been ranked as a leading researcher in Group Support Systems. "

Del Hawley is the Senior Associate Dean and Associate Professor of Finance at the School of Business Administration at The University of Mississippi. He holds a Ph.D. and MBA in Finance and a B.S. in Psychology from Michigan State University, and has more than 20 years of academic experience and 15 years of prior experience in business management. In his administrative role, he has served as the CFO, COO, and CIO for the business school for 15 years.

Mahesh Vanjani is a Professor of Management Information Systems in the Jesse H. Jones School of Business at Texas Southern University. He has published numerous articles and made several presentation based on his research. His research interests include group decision support systems, group behavior, electronic language translation, e-commerce, e-government and Internet use.