7-1-2006

Implications Based Upon Preliminary Analyses of a Student Evaluation of Teaching Database

Jack Yurkiewicz
Pace University

Peter Hoefer
Pace University

John Byrne
Pace University

Working Papers

No. 221 July 2006

IMPLICATIONS BASED UPON PRELIMINARY ANALYSES OF A STUDENT EVALUATION OF TEACHING DATABASE

by

Jack Yurkiewicz, Ph.D.
Professor of Management and Management Science
Lubin School of Business
Pace University

Peter Hoefer, Ph.D.
Professor of Management and Management Science
Lubin School of Business

and

John Byrne, Ph.D.
Assistant Professor of Management and Management Science
Lubin School of Business

IMPLICATIONS BASED UPON PRELIMINARY ANALYSES OF A STUDENT EVALUATION OF TEACHING DATABASE

by

Jack Yurkiewicz, Ph.D.

Peter Hoefer, Ph.D.

and

John Byrne, Ph.D.

Jack Yurkiewicz is Professor of Management and Management Science at the Lubin School of Business, Pace University.

Peter Hoefer is Professor of Management and Management Science at the Lubin School of Business, Pace University.

John Byrne is Assistant Professor of Management and Management Science at the Lubin School of Business, Pace University.

ABSTRACT

Student evaluations of teaching data are analyzed for a semester of classes at a large collegiate business school. Summary statistics and other measures show that biases are evident in the overall database, suggesting that caution is needed if the data is to be used in any decision-making capacity.

1. INTRODUCTION

Student evaluations of teaching (we will use the popular "SET" abbreviation) instruments have been present at the Lubin School of Business at Pace University since the 1970s. The current instrument was introduced in 1991, following approval by the Lubin Graduate School Faculty Council (now, the Lubin Faculty Council). The instrument is currently intended to be administered to every non-independent-study Lubin class during the Fall and Spring semesters, and to classes by request during the two summer semesters. Administration of the evaluations is performed during the last few weeks of the semester. Validated questionnaires (that is, questionnaires that the administration believes are legitimate) are summarized electronically. The electronic summaries are provided to all stakeholders about six to eight weeks after the end of the semester, ensuring grades are submitted before the summaries are made available. The original questionnaires are given to the faculty member only.

As with many instruments being used at other institutions (see section 2), the instrument is a compromise document which is intended to serve many constituencies. In the late 1990s, a Lubin School Faculty Council committee was commissioned to review the instrument and the process, and they considered modifying the questionnaire. However, a modified questionnaire was not approved by the Council, so the original questionnaire is still used today.

The purpose of our studying a large database associated with the SET is to understand the strengths and weaknesses inherent in the SET and its use in other processes (for example, the annual faculty evaluation process). The initial paper discusses the instrument, looks at basic implications, and suggests further analysis and research.

2. PRIOR RESEARCH

The analysis of SETs is usually concentrated either on the instrument itself, or on how the results of the SET are related to some other variable. For example, much of what we know about the relationship between SETs and student "performance" is directly, or indirectly, related to Peter Cohen's (1981) meta-analysis of the relationships between SETs and student achievement. That study analyzed the results associated with 41 independent validity studies related to our topic of interest. It sets forth expectations about validity studies associated with the relationship between student evaluations and student achievement. Its conclusion offers strong support for the validity of student evaluations as measures of teaching effectiveness.

A continuation of Cohen's study was performed and reported by Feldman (1989). In that paper, Feldman began with the same data set as Cohen, but the relationships were refined. A major result is that the information analyzed by Feldman involved the relationship between specific instructional dimensions and student achievement, rather than overall teaching evaluation and achievement.

Krautmann and Sander (1997) investigated the relationship between SETs and a student's expected grades based on empirical data gleaned from economics courses at DePaul University. They paid specific attention to the possibility of expected grades being an endogenous variable, which would render ordinary least squares regression equations biased. Even though they found no evidence of endogeneity, they used two models, ordinary least squares and two-step least squares. In both situations, they found a significant relationship between the SET and expected grades. In Isely and Singh (2005), the relationship between a student's expected grades, previously attained GPA, and the SET were analyzed. They found that the difference between expected grade and GPA affected the SET significantly. Stapleton and Murkison (2001), in a study of 29 faculty members in their department at a large southern university, included an analysis of the relationship between SETs and expected grades. They found a strong relationship between the two variables. Whitworth, Price, and Randall (2002) investigated a very large set of management classes offered at a southeastern university over a three-year period. They found a significant difference in SET scores when compared with students' perception of learning. They also found that gender, academic level (graduate or undergraduate), and course subject played a role in determining SET scores, suggesting that using SETs to rank faculty as a whole is not a valid procedure. They used the standard difference between means and supported that analysis with a Chi-square analysis of the associated contingency table.

In Yunker and Yunker (2003), a *negative* relationship was discovered between controlled student learning and average SETs for faculty in multiple sections of a core accounting course. The results are different from most other broader-based results, but they do involve the use of average SETs, which appear to be more appropriate, and are what we use in our study. The implications are severe. Seiler and Seiler (2002) performed a structural equation model analysis of data obtained from accounting courses at a medium sized Midwest university. The analysis of the 500 plus student evaluations indicated a significant relationship between "perceived student learning" and "overall instructor evaluation."

Other studies concern themselves with different issues associated with the SET process. Williams and Ceci's (1997) report on the findings of a Cornell experiment is one of particular note. In that experiment, a professor taught a course in one semester, and taught the identical course the next semester with one change: He used a much more enthusiastic voice throughout the second semester. The findings report that the SETs improved in every dimension in the more enthusiastically-taught course.

Green, Calderone, and Reider (1998) performed a content analysis of teaching instruments used in accounting departments at the collegiate level. They recommended that most accounting departments needed to redesign their instruments to include information on the dimensions of teaching that students could actually assess. They also suggested that if these instruments were used in an overall faculty assessment process, an overall portfolio should be used instead of just the SETs. Using MIS courses, Simon and Soliman (2003) approached SETs specific to that area of study. They concentrated on measuring students' attitudes towards, and perceptions of, subject matter, and

recommended a modified approach for evaluation of faculty. McKeachie (1987) looked at improving the overall evaluation process, with suggestions about specific measures in the SET and how to apply the SET to improve teaching. Marsh and Roche (1997) stressed the need to identify the components of teaching in SETs. Jacobs and Kozlowski (1985) studied the "halo effect" (positive & negative) arising from raters' familiarity with the subject they are rating. In this context, students taking a faculty member for multiple courses are obviously familiar with the instructor they are rating in a SET. Boex (2000) studied the characteristics of perceived good teaching of economics at Georgia State University, using a SET he developed. He determined that two dimensions stood out strongly, namely organization and clarity of the instructors and teaching material. Dunegan and Hrivnak (2003) questioned the validity of the SET by focusing on whether or not students actually paid attention to completing the form. Marks (2000) raises questions about discriminant validity and again questions the uses of summary measures employed to describe the faculty member.

McKone (1999) analyzed data associated with a SET which existed at the Darden School at the University of Virginia. The SET had 29 questions; she identified relationships that depended, as usual, on two main factors (course and faculty) and proceeded to provide a model for feedback to instructors who sought to increase their ratings, as well as to administrators who sought a fair evaluation of faculty. A result was a simpler, 14 question instrument.

Morgan, Sneed, and Swinney (2003) surveyed accounting faculty and University administrators to seek similarities and differences in perceptions concerning SETs. Not surprisingly, they found that administrators believed SETs measure teaching effectiveness to a greater degree than faculty, who believed their personalities carried the most weight on SETs.

Campbell, Gerdes, and Steiner (2005) concerned their study with how the "attractiveness" of an instructor relates to SETs. Whereas there were a number of studies done in the past that indicated a significant correlation between attractiveness and SETs, their hypothesis was that the positive findings were biased because of omitted variables. In their study, they controlled for other variables and found no significant relationship between attractiveness and SETs.

A summary of the prior research may be best described by stating that much was studied, and, depending upon the type of database, analysis, and control, many hypotheses were supported and unsupported. However, the research we are reporting on is specific to recent data associated specifically with the Lubin School, and implications for the Lubin School are important.

3. ANALYSIS AND INTERPRETATION

3a. Overall Data Description and Analysis

There have been numerous student evaluations of faculty (SETs) performed at the Lubin School of Business at Pace University since the 1980s. Figure 1 contains the closed-ended questions from the current instrument, which was adopted by the Lubin Faculty Council in 1991 and implemented by the Lubin administration shortly thereafter. There are 10 closed answer questions, followed by open answer questions. Each of the 10 closed answer questions form a 5-point Likert-scale, where 1 is "worst" and 5 is "best." The raw data is only available to the faculty member, who can share the information with whomever she or he decides. The summary data for the 10 closed answer questions are "public," in the sense that they are available to Deans, Chairs, and students.

Figure 1

1	How worthwhile did you find this course?	Not at all worthwhile	1	2	3	4	5	Very worthwhile
2	The professor explained the course requirements	Not at all clearly	1	2	3	4	5	Very clearly
3	The professor explained the grading system	Not at all clearly	1	2	3	4	5	Very clearly
4	How would you rate the professor's preparation for class sessions?	Poorly prepared	1	2	3	4	5	Very well prepared
5	The professor's attitude towards the course subject matter is	Negative	1	2	3	4	5	Enthusiastic
6	How satisfied were you with the professor's availability?	Not at all	1	2	3	4	5	Completely
7	How satisfied were you with the professor's respect for students?	Not at all	1	2	3	4	5	Completely
8	How satisfied were you with the professor's management of classroom time?	Not at all	1	2	3	4	5	Completely
9	How satisfied were you with feedback provided by the professor throughout this course?	Not at all	1	2	3	4	5	Completely
10	Overall, how would you rate the professor?	Poor	1	2	3	4	5	Outstanding

Each semester, the goal of the administration is to survey all business courses. In practice, about 95 percent or more of the non-independent study courses are evaluated. We began with a list of all the courses in the Fall 2004 semester for which we had evaluations. We then added to the 10 summarized SET questions other pertinent available information about the course or instructor.

When looking at the SET information, we were cognizant that all of the questions except question 1 directly refer to the faculty member leading the course. We also noted that perhaps the most important question in that group relative to the purpose of this analysis is question 10. That question has been used by administrators specifically during the annual faculty teaching evaluation process and has therefore taken on a role of importance with faculty, administration, and students. A factor analysis of the 10 questions confirms the role of question 10. The analysis of the questionnaire produced one component for all questions, using the latent root criterion with an eigenvalue above 1 for extraction. The one component explained 82 percent of the variance in the questionnaire and loaded most highly on question 10 (.975). We validated the factor analysis by splitting the database, again achieving the same single factor with similar loadings. For that reason, we use question 10 as a surrogate for the one factor questionnaire.

Along with the 10 summarized SET questions, the database also consists of the course reference number, the course, and department code (ACC320, for example), the department, the level (graduate or undergraduate), the gender of the professor, the grade distribution for the course, the average QPA for the course, the course enrollment as of October 15, 2004, and the number of students completing the questionnaire. A preliminary analysis gave a small, insignificant difference between the number of students who didn't complete the survey and the question 10 averages. The concern was that those students who did not complete the survey could have a significant impact on the overall evaluation.

We then "cleansed" the data by removing courses that were considered inappropriate to this analysis (for example, a course that had three students in it that was not listed as an independent study course). We also did not include executive MBA or doctoral courses, as their delivery and expectations we considered to be significantly different from the listed BBA, MBA, and MS courses. The end result was 381 courses for which we were comfortable that the information we had was reliable.

Preliminary analysis suggested the scaling differences between undergraduate and graduate course grades was causing difficulty, so we "normalized" the graduate QPA on a 1-4 scale, the same scale used by the undergraduate courses (the graduate scale was 1-3, or A-C, followed by F). This was done by a simple linear transformation of the average graduate course grades, only. When grades are discussed, most of the following analysis is based on the normalized QPA scale, except where noted.

Looking at some of the key variables, assuming they can be summarized, provides us with the following:

Table 1: Overall Summaries

	Mean Grade	Mean Grade Normalized	Q1	Q2	Q3	Q4
Mean	3.1520	3.0598	4.2024	4.3790	4.4178	4.4604
Median	3.2000	3.1079	4.2300	4.4800	4.5000	4.5800
Std. Dev.	0.4384	0.4224	0.4537	0.4300	0.3958	0.4432
Count	381	381	381	381	381	381

	Q5	Q6	Q7	Q8	Q9	Q10
Mean	4.5633	4.3233	4.4890	4.3186	4.2365	4.2977
Median	4.6400	4.4000	4.5900	4.4300	4.3200	4.3800
Std. Dev.	0.3672	0.4397	0.4469	0.5192	0.4865	0.4955
Count	381	381	381	381	381	381

All the question distributions show a rather high average. However, the questionnaire is "mature," and some may infer that the faculty have learned to "teach to the questionnaire."

This analysis will concentrate on the faculty member rather than the student. Therefore, we will compare *averages* of class scores, rather than the individual student evaluations. There are arguments for and against such a procedure. A prominent argument for using averages in this context is presented in Linn, Centra, and Tucker (1975). The essence of the argument is that we are analyzing differences in faculty, not students.

The simple Pearson correlation between the average of question 10 and the normalized student average grades is .187. The p-value for this correlation is a strong .000244, in part due to the reasonably large sample size. The significant positive correlation is consistent with, but lower than, the results found in other studies. It is on the low end of Cohen's average of multisection-course correlation between overall instructor ratings and student achievement, which was found to be .43 overall (Cohen, 1981). Another comparative correlation is in Stapleton and Murkison (2001), where the correlation between two similar questions about instructor excellence and expected grade was .26.

However, as we will see later, our low findings are perhaps biases due to the relationship between other variables and the SET (gender, level, and type of course).

Using a contingency table analysis and the Chi-square statistic, with the categorized average of question 10 forming the rows and normalized grades forming the columns, we get a similar result indicating rejection of the null hypothesis of independence (that is, question 10 and the normalized grades are dependent) with a p-value of .01001. Using Spearman's rank-correlation gives an r of .193, which is also significant at the .01 level of significance, and similar to the simple correlation.

3b. Gender Specific Segmentation

When seeking to understand possible gender differences in the relationship between average SET scores and average course grades, there are a number of striking contrasts. First, consistent with business schools, the number of courses taught by men at the Lubin School is significantly larger than those taught be women. The number of female-taught sections numbers 59, whereas the number of male-taught sections numbers 322.

Examining the normalized average course grades, female-taught sections averaged 3.0630 and male-taught sections averaged 3.0592, practically the same (literally) and not significantly different in a statistical sense. In comparing the question 10 average responses, female-taught sections averaged 4.4203, whereas the male-taught sections averaged 4.2742, significant at the 5 percent level. This study showed that female professors scored higher than male professors.

Still looking at the gender issue, the variables gender and question 10 average are statistically dependent at the 5 percent level, using a contingency table analysis and the Chi-square statistic. Using simple correlation, the dummy variable gender correlates with question 10 average with an r of -.106, significant at 5 percent. This supports the conclusion that as we switch from female to male faculty, the average question 10 score will decrease, consistent with the data above.

Looking at the issue separately, that is, looking at female data and then male data, also leads to some striking contrasts. The correlation between women average student grades and question 10 is .279; this is statistically significant at 5 percent. For men, the corresponding correlation is .176; this is statistically significant at 1 percent, probably because of the large sample size (322).

We conclude that female-taught courses have higher SET question 10 scores than male-taught courses. Also, there is a stronger relationship between average student grades and SET scores for women than for men. Overall, in this context, there are interesting gender-specific differences in the data.

3c. Academic Level Specific Segmentation

The average graduate grade assigned is 3.48, the average undergraduate grade assigned is 2.97, and the normalized graduate grade is 3.22. There is a significant difference among these three numbers, as well as pair-wise differences; all are at the 1 percent level.

When comparing the SETs by level, there are again significant differences between the undergraduate SETs (4.35), which are significantly higher, and the graduate SETs (4.21). The level of significance is at 1 percent. A contingency table Chi-square test confirms the dependence of SET scores on level (graduate and undergraduate), the result again being at the 1 percent level of significance.

The segregation of the data into graduate and undergraduate databases allows for higher correlations between the student grades and the SETs. For graduate students, the correlation between average student grades and average question 10 is .226 (significant at 1 percent level). Note that the correlation is the same for the normalized grades in the graduate segment. For undergraduate students, the correlation between average student grades and average question 10 is .243 (significant at the 1 percent level). These results are confirmed by contingency table analyses using the Chi-square test.

We conclude that the faculty teaching undergraduate courses scored significantly higher on the average of question 10 than the faculty teaching graduate courses. We also observe a significant difference in assigned and normalized grades, probably due to the difference in scale and process as indicated in an earlier section. Finally, both graduate courses and undergraduate courses show a relatively high, significant correlation between average student grades and question 10 averages. The suggestion is that the segmentation of the data is appropriate for further analysis, and the correlations have increased because of that segmentation.

3d: Department Specific Segmentation: Accounting Department

Table 2: Accounting Department Summaries

	Mean Grade	Mean Grade Normalized	Q1	Q2	Q3	Q4
Mean	2.8730	2.8103	4.2031	4.4224	4.4944	4.5135
Median	2.7850	2.7706	4.2100	4.4400	4.5100	4.5700
Std. Dev.	0.4724	0.4296	0.3793	0.3488	0.3449	0.3730
Count	75	75	75	75	75	75

	Q5	Q6	Q7	Q8	Q9	Q10
Mean	4.5324	4.3227	4.4843	4.3532	4.2705	4.3156
Median	4.6000	4.3300	4.5900	4.4400	4.3100	4.3800
Std. Dev.	0.3277	0.3728	0.4437	0.6172	0.4438	0.4208
Count	75	75	75	75	75	75

For each of the departments, we feel it is worthwhile to present summary measures associated with the SET, as well as average grades. The Accounting Department sample shows the lowest average grades (by either measure). It also shows an approximate average response to questions 1, 6, and 7; above average response to questions 2, 3, 4, 8, 9 and 10; and below average response to question 5.

The simple correlation between the normalized grades and (the surrogate) question 10 is .097 ($p = .406$).

3e: Department Specific Segmentation: Finance and Economics Department

Table 3: Finance and Economics Department Summaries

	Mean Grade	Mean Grade Normalized	Q1	Q2	Q3	Q4
Mean	3.3577	3.2176	4.2392	4.3129	4.4042	4.3867
Median	3.4333	3.3037	4.3300	4.3900	4.5000	4.5200
Std. Dev.	0.3565	0.4356	0.4097	0.4837	0.3913	0.5123
Count	76	76	76	76	76	76
	Q5	Q6	Q7	Q8	Q9	Q10
Mean	4.5407	4.2721	4.4447	4.2353	4.1609	4.2245
Median	4.5800	4.4000	4.6000	4.3200	4.2700	4.3200
Std. Dev.	0.3521	0.4625	0.5187	0.5621	0.5260	0.5376
Count	76	76	76	76	76	76

The Finance and Economics Department sample shows the highest average grades (by either measure). It also shows an approximate above average response to question 1; and below average response to questions 2, 3, 4, 5, 6, 7, 8, 9, and 10. Recall question 1 refers to the course.

The simple correlation between the normalized grades and (the surrogate) question 10 is .113 (p=.330).

3f: Department Specific Segmentation: Legal Studies and Taxation Department

Table 4: Legal Studies and Taxation Department Summaries

	Mean Grade	Mean Grade Normalized	Q1	Q2	Q3	Q4
Mean	3.0435	2.9336	4.3700	4.4305	4.4155	4.5404
Median	3.0385	2.9300	4.4000	4.5000	4.4800	4.5600
Std. Dev.	0.3917	0.3875	0.3604	0.3567	0.3274	0.3441
Count	56	56	56	56	56	56
	Q5	Q6	Q7	Q8	Q9	Q10
Mean	4.5798	4.3852	4.5321	4.3916	4.2884	4.3946
Median	4.6200	4.4600	4.5500	4.4500	4.3300	4.5000
Std. Dev.	0.3292	0.3541	0.3551	0.3837	0.4054	0.4129
Count	56	56	56	56	56	56

The Legal Studies and Taxation Department sample shows below average mean grades. It shows an approximate average response to question 3; it shows an above average response to questions 1, 2, 5, 6, 7, 8, 9, and 10; and below average response to question 4.

The simple correlation between the normalized grades and (the surrogate) question 10 is .192 (p=.156).

3g: Department Specific Segmentation: Marketing Department

Table 5: Marketing Department Summaries

	Mean Grade	Mean Grade Normalized	Q1	Q2	Q3	Q4
Mean	3.1783	3.1153	4.2529	4.4314	4.4057	4.5088
Median	3.1789	3.1405	4.3000	4.5400	4.5200	4.5800
Std. Dev.	0.4050	0.3709	0.4212	0.4484	0.4676	0.3942
Count	56	56	56	56	56	56
	Q5	Q6	Q7	Q8	Q9	Q10
Mean	4.6845	4.4041	4.5675	4.3839	4.3250	4.4066
Median	4.7400	4.4800	4.6700	4.5000	4.4700	4.5000
Std. Dev.	0.2828	0.4129	0.3415	0.4265	0.4512	0.4564
Count	56	56	56	56	56	56

The Marketing Department sample shows above average mean grades. It shows an above average response to questions 2, 4, 5, 7, 8, 9, and 10; and below average response to questions 1, 3, 6.

The simple correlation between the normalized grades and (the surrogate) question 10 is .353 (significant at the .01 level, p=.008).

3h: Department Specific Segmentation: Management and Management Science Department

Table 6: Management and Management Science Department Summaries

	Mean Grade	Mean Grade Normalized	Q1	Q2	Q3	Q4
Mean	3.2358	3.1501	4.0747	4.3445	4.3846	4.4134
Median	3.3200	3.1857	4.1200	4.4800	4.4700	4.5900
Std. Dev.	0.4024	0.3640	0.5417	0.4599	0.4209	0.4911
Count	118	118	118	118	118	118
	Q5	Q6	Q7	Q8	Q9	Q10
Mean	4.5322	4.2889	4.4627	4.2847	4.1969	4.2358
Median	4.5900	4.3600	4.5500	4.4200	4.2900	4.3200
Std. Dev.	0.4392	0.5059	0.4814	0.5160	0.5318	0.5516
Count	118	118	118	118	118	118

The Management and Management Science Department sample shows above average mean grades. It shows a below average response to questions 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 (all questions).

The simple correlation between the normalized grades and (the surrogate) question 10 is .373 (significant at the .01 level, p=.000).

4. DISCUSSION AND RECOMMENDATIONS FOR CONTINUING RESEARCH

In section 3a, when looking at the entire data set, we noted a significant, but relatively low, positive correlation between the surrogate question 10 and average grades assigned for a course. After perusing the results from section 3c onward, it is possible that the low correlation (as compared with other studies; again refer to Cohen 1981) occurs because of essentially different data sets, that is, data sets that are exhibiting a different structure. This is especially true when comparing academic level, graduate versus undergraduate, even after normalizing the grade distributions. An obvious suggestion for further research is to report on these two data sets individually.

In section 3b, we noted that even though the average grades are approximately the same for men and women, women averaged higher for question 10. We also noted a "penalty" for being a male teacher in a regression model with a dummy gender variable.

As suggested in the first paragraph above, perhaps the most important suggestion from section 3c, academic level, is that it appears as if the two data sets are different. Segregating the database into graduate and undergraduate levels causes the correlations to become more in line with what was found in other studies (Cohen 1981, and others). Also, as was indicated at the end of 3c, graduate grades are significantly higher than undergraduate grades, and undergraduate SETs are significantly higher than graduate SETs (using the surrogate question 10 as representative of the SET). Whereas differences in requirements for graduation may appropriately rationalize the difference in grades (undergraduate students need to maintain a 2.00 QPA to graduate, whereas graduate students need to maintain a 3.00 QPA), the significant difference in question 10 average has severe implications.

Finally, in sections 3d-3h, the differences in SETs among departments suggests that one must be careful if using the SET in a singular evaluative procedure for the entire faculty. There are numerous uncontrollable variables that may cause these differences. For example, the type of material taught (quantitative versus qualitative), classroom methodology (lecture versus case study with groups), may be having an effect on the differences in SETs.

There is much room for further analysis of this data set. Again, we will only provide a partial list. Segregating by gender and level and department is important. Even within broad department categories, additional segregation (for example, as management science faculty, we are naturally interested in separating this section out from the other topics in the Management and Management Science Department) could be examined. Work has already begun on some of the topics we have indicated, and we seek comments to improve the quality and accuracy of this report.

BIBLIOGRAPHY

Boex, L.F. Jameson. "Attributes of Effective Economics Instructors: An Analysis of Student Evaluation." *Journal of Economic Education* 31, no. 3 (Summer 2000): 211-227.

Campbell, Heather E., Karen Gerdes, and Sue Steiner. "What's Looks Got to Do With It? Instructor Appearance and Student Evaluations of Teaching." *Journal of Policy Analysis and Management* 24, no. 3 (2005): 611-620.

Cohen, Peter A. "Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies." *Review of Educational Research* 51, no. 3 (Fall 1981): 281-309.

Dunegan, Kenneth J., and Mary W. Hrivnak. "Characteristics of Mindless Teaching Evaluations and the Moderating Effects of Image Compatibility." *Journal of Management Education* 27, no. 3 (June 2003): 280-303.

Feldman, Kenneth A. "The Association Between Student Ratings of Specific Instructional Dimensions and Student Achievement: Refining and Extending the Synthesis of Data from Multisection Validity Studies." *Research in Higher Education* 30, no. 6 (1989): 583-645.

Green, Brian Patrick, Thomas G. Calderone, and Barbara Powell. "A Content Analysis of Teaching Evaluation Instruments Used in Accounting Departments." *Issues in Accounting Education* 13, no. 1 (February 1998): 15-30.

Isely, Paul, and Harinder Singh. "Do Higher Grades Lead to Higher Student Evaluations?" *Journal of Economic Education* 36, no. 1 (Winter 2005): 29-32.

Jacobs, Rick, and Steve W. J. Kozlowski. "A Closer Look at Halo Error in Performance Ratings." *Academy of Management Journal* 25, no. 1 (1985): 201-212.

Krautmann, Anthony C., and William Sander. "Grades and Student Evaluations of Teachers." *Economics of Education Review* 18 (1999): 59-63.

Linn, Robert L., John A. Centra, and Ledyard Tucker. "Between, Within and Total Factor Analyses of Student Ratings of Instruction." *Multivariate Behavioral Research* XX, no. XX (July 1975): 277-288.

Marks, Ronald B. "Determinants of Student Evaluations of Global Measures of Instructor and Course Value." *Journal of Marketing Education* 22, no. 2 (August 2000): 108-119.

Marsh, Herbert W., and Lawrence A. Roche. "Making Students Evaluation of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias and Utility." *American Psychologist* 52, no. 11 (November 1997): 1187-1197.

McKeachie, Wilbert J. "Instructional Evaluation: Current Issues and Possible Improvements." *Journal of Higher Education* 58, no. 3 (May/June 1987): 344-350.

McKone, Kathleen E. "Analysis of Student Feedback Improves Instructor Effectiveness." *Journal of Management Education* 23, no. 4 (August 1999): 396-415.

Morgan, Donald Ace, John Sneed, and Laurie Swinney. "Are Student Evaluations a Valid Measure of Teaching Effectiveness? Perceptions of Accounting Faculty Members and Administrators." *Management Research News* 26, no. 7 (2003): 17-32.

Simon, Judith C., and Khalid S. Soliman. "An Alternative Method to Measure MIS Faculty Teaching Performance." *The International Journal of Education Management* 17, nos.4/5 (2003): 195-199.

Stapleton, Richard John, and Gene Murkison. "Optimizing the Fairness of Student Evaluations: A Study of the Correlations Between Instructor Excellence, Study Production, Learning Production, and Expected Grades." *Journal of Management Education* 25, no. 3 (June 2001): 269-291.

Whitworth, James E., Barbara A. Price, and Cindy H. Randall. "Factors That Affect College of Business Student Opinion of Teaching and Learning." *Journal of Education for Business* (May/June 2002): 282-289.

Williams, Wendy M., and Stephen J. Ceci. "How'm I Doing?" *Change* 29, no. 5 (September/October 1997): 13-24.

Yunker, Penelope J., and James A. Yunker. "Are Student Evaluations of Teaching Valid? Evidence From an Analytical Business Core Course." *Journal of Education for Business* 78, no. 6 (July/August 2003): 313-317.