Students Evaluations: What is a Chair to Do?
Preliminary Data Analysis

William P. O'Dea*

ABSTRACT

The purpose of this project is to develop a set of metrics that department chairs can use to evaluate the results of their student evaluations. In this phase of the project, I consider three issues: are the rankings of instructors stable over time; are the rankings influenced by factors beyond the control of the instructor such as the level of the class; and are the rankings influenced by the instructor's grading policy. Preliminary research indicates that the answers are respectively: yes, yes, and no. The next step in the project will be to determine whether the student evaluations can identify the qualities possessed by effective teachers.

I. INTRODUCTION

I am an end-user of student evaluation of instruction (SEI) data. In my role as department chair, I write letters of recommendation for faculty members requesting contract renewals, tenure or promotion. In these letters, I am expected to assess their teaching performance. We ask our faculty to provide a portfolio of material to document their teaching performance. In addition to summaries of the SEI data, this portfolio includes syllabi, copies of examinations, peer reviews based on classroom visits, and a self-evaluation. The SEI data tend to attract the lion's share of the attention in faculty discussions because they come in an easily digestible form and give at least the illusion of precision. I am also a member of my college's Promotion and Tenure Committee (P&TC). Attempts to parse the meaning of the SEI data play a significant role in our discussions of teaching effectiveness.

Ideally, we would like the SEI data to tell us something about an instructor's ability to convey a body of knowledge and train students to apply concepts developed in their courses to unstructured problems. Of course, we would also like students to retain what they learn.

At my institution, many members of the faculty are not convinced that the SEI data provide any valuable insights into a faculty member's teaching effectiveness. Some of my fellow department chairs have expressed the opinion that the SEI results are pure noise and are determined by the academic equivalent of the spin of a roulette wheel. One of my colleagues on P&TC was convinced that the SEI data did not measure teaching effectiveness but rather provided a venue in which our students could vent their racist, sexist, homophobic and xenophobic attitudes. In this vein, the Marlin/Niss Principle of Teacher evaluations is worthy of note. The Principle holds that: "Teachers who have received high student evaluations in the past

*SUNY Oneonta, Department of Economics, Finance and Accounting – Email: odeawp@oneonta.edu

will find them to be valuable measures of good teaching. Teachers who have received low students evaluations in the past will find them to be laughably insignificant (Marlin and Niss 1980, p. 25)".

In any assessment of the value of the SEI data, a good starting point is the literature on the subject, which is vast.[1] As of 1999, Becker and Watts (1999) report that "nearly 2000 studies have been conducted (p. 344)." Since 1978, over twenty articles on student evaluations have been published in the *Journal of Economic Education.* By any standards, the intellectual effort that has gone into evaluating the SEI data is impressive. The literature has three major concerns. The first is to identify the important explanatory variables that drive the SEI results. The second is to establish the relation between grading standards and student evaluations. Siegfried and Fels (1979) put the matter bluntly: "By far the most controversial issue is whether instructors can 'buy' higher evaluations by lowering the (effort) price to students of achieving a given grade (p. 931)." The third focus of the literature is to identify proper estimation techniques. While ordinary least squares (OLS) is the low-cost solution to any estimation problem, it may not be the ideal approach to employ in the analysis of SEI data. Given my role as a consumer of SEI data, I am going to focus on the first two concerns.

With regard to the question of what instructor qualities are valued by students, the literature is reassuring. Boex (2000) reports that "from the students' point of view, organization and clarity was the single most important attribute of effective economics instruction (p. 213)." Bosshardt and Watts (2001) find that students care about enthusiasm and preparation. DeCanio's (1986) results highlight the importance of organization and structure. Nelson and Lynch (1984) find "clarity of communication and instructor enthusiasm" to be the most important factors explaining favorable student evaluations (p. 21).

Concerning the impact of grades on evaluations, the results in the literature are mixed. Some authors such as DeCanio (1986) find "no evidence that the expected grade influences the instructor's teaching effectiveness score (p. 172)." Other authors find that the impact is small. For example, Nelson and Lynch (1984) report that a one point increase in the average grade in a course (on a four point scale) only raises the instructor rating by .15 (on a five point scale). McPherson (2006) finds a more pronounced effect: a one point increase in the average expected grade raises the teaching effectiveness score by .34.

Of course, establishing the impact of grades on the SEI ratings is not easy. Higher grades are not necessarily an indication of lax grading standards or a naked attempt to curry students favor. Higher grades could be the result of effective instruction that results in a high level of student learning. Ideally, we would like to hold the level of student learning constant in order to discover whether instructors with more lenient grading standards receive higher evaluations. The problem is that since examination design and grading are generally under the control of individual instructors it is difficult to find an independent measure of what students have learned. This is not to say that attempts have not been made. For example, Weinberg et al. (2009) use student performance in follow-up courses as a learning measure. Depressingly, their results show that "student evaluations are strongly related to grades and that learning, as measured by future grades, is unrelated to student evaluations once current grades have been controlled (p. 254). " Of course, it is also important to hold more than learning constant. Isley and Singh (2005) suggest that self-selection on

the part of students is a confounding factor. Since students know the identity of their course instructors in advance and have access to information sources such as R*ate My Professor*, it is possible that some instructors have developed reputations that attract hard working, highly motivated students and repel the casually dedicated. The result would be a high level of student performance that translates into high grades and high SEI ratings. Their results suggest that such a causal factor is at work.

The fact that the literature despite its bulk offers relatively few hard results is not that surprising. Student evaluation instruments are extremely heterogeneous. I have looked at thirteen instruments from other institutions. While this sample is probably not representative, a few observations are in order. The first is the instruments vary widely in length from a relatively compact ten questions at Ohio State (Weinberg et al. 2009) to the 50 question workout used at Ball State (Dilts 1980). The surveys generally cover the same basic concerns: organization and delivery of lectures; course management; examination design and grading; and classroom demeanor. Some of the surveys have "bottom line" summative questions in which the students are asked to rate the value of the course or the instructor's overall performance. For example, on our instrument, the last question asks for "an overall evaluation of teaching effectiveness." Other instruments do not. The presence of a summative question is important since by its nature this type of question tends to attract more attention than the other questions in the instrument.

Even if a uniform evaluation instrument were in use, there is no reason to expect that analysis of the data would yield equally uniform results. Institutions differ with regard to factors such as the emphasis they place on teaching versus research, whether they are populated by residents or commuters, and the age of the student body. Even at the same institution, it is likely that the SEI ratings could differ across disciplines. Becker (1997) notes that "economics is one of the disciplines that is consistently at the bottom of both course and instructor effectiveness scales… (p. 1369)."

As a department chair, I am in the position of having to make sense of the data that are generated by my department's evaluation instrument. I did not design the instrument. Owing to faculty inertia and/or risk aversion, changing the instrument would be difficult. I have no access to student level data. All I receive are the statistical summaries of the results for each course section. The lack of student level data is not necessarily a bad thing. Weinberg et al. argue that if the goal is to establish the connection between grades and teaching ratings, it is entirely appropriate to focus on class-level data. I have no independent measure of what students have learned in a course. In the literature, measures of student learning tend to be based on either a comparison of pre-and post-test results (for example, Marlin and Niss 1980) or, as noted above, an analysis of student performance in follow-up courses. I have no great interest in pursuing either approach. The first would make a sizable dent in my budget and try the patience of my faculty. The second would be difficult to implement and would represent a major addition to my workload. I have access to the distribution of grades in each course section. What I do not have access to are the courses grades (or the expected course grades) of the subset of students in each course who actually complete the questionnaires.

In short, I am the proprietor of an ongoing experiment that was not well-designed. Consequently, I have at my disposal an imperfect data set. My task is to extract the maximum possible meaning from this data, assuming that there is meaning to be extracted.

This paper represents my preliminary analysis of my data. I would like to answer three questions. First, are the data so noisy as to be essentially meaningless? Second, to what extent are the results influenced by factors beyond the control of the faculty members? In addition to the alleged antisocial tendencies attributed to our students by my colleague on P&TC, these factors could include the level of the course, the number of students enrolled in the course, and the time of day at which it is offered. Third, to what extent are the SEI results influenced by factors under the control of the faculty members such as their grading standards and course design and delivery. My ultimate objection is to develop a set of simple metrics that can be used to answer these questions and a means of adjusting the SEI results to account for factors beyond a faculty member's control.

II. BACKGROUND AND SUMMARY DATA

At Oneonta, it is College policy that the SEI questionnaires be administered in every course every semester. Individual departments can use a College questionnaire or substitute an instrument of their own design. The Division of Economics and Business, which houses the Departments of Economics, Finance and Accounting (EFA) and Management, Marketing and Information Systems, uses its own questionnaire.

The timing and mechanics of administering the SEI instrument are up to the individual faculty member. Typically, the SEIs are administered during the last two weeks of the semester. Some brave souls administer the evaluations during the final examination period. Instructors are not supposed to administer the questionnaires. Some instructors recruit other members of the faculty to hand out and collect the forms. Other instructors explain the process, distribute the forms, and designate a student in the class to collect the forms and return them to the department office. The department secretary then types up the written comments. The forms are then sent to the Office of Institutional Research which generates summary statistics and provides analysis of the individual questions.

The questionnaire used by the Department of Economics, Finance and Accounting (EFA) is relatively compact. It asks the students to answer thirteen questions and provide their GPAs. Unlike the instruments at other institutions, we don't ask the students to provide their expected grade for the course or to rate the value of the course. Our bottom line question focuses on instructor effectiveness. On our form, one is excellent (or the equivalent) and five is poor. On the back of the form, the students can provide written comments. The student comments tend to be terse. If our students write more than a sentence or two, it is generally a sign that they are not pleased.

Table 1 provides data on the average scores for the thirteen questions for the course sections taught in EFA from the fall of 2008 to the spring of 2012. There were a total of 277 course sections taught by nineteen different instructors. In general, the scores are slightly over two which means that on the" overall evaluation

of teaching effectiveness" and most of the individual components of instructor performance our students think their instructors are (almost) very good.

It is also possible that course and instructor characteristics affect the student evaluations. Table 2 presents summary data for course and instructor characteristics that have been employed as explanatory variables in other studies. The number of students enrolled in the course captures the possibility that students might prefer smaller classes because they allow more interaction between the instructor and the students. The percentage of surviving students who complete the SEI form accounts for the possibility that the sample of students completing the SEI forms might not be representative of the entire population of students in the course. Thus, the SEI data are subject to sample selection bias. Boex (2000) raises the possibility that students who are disappointed with a course are likely to stop attending class and thus will not fill out the SEI questionnaire. The result would be a higher overall SEI rating. Becker and Watts (1999) suggest that instructors might manipulate the response rate by deciding when to administer the form. For example, at our institution, an instructor might administer the evaluation on a Friday, which is a day with a relatively high absenteeism rate. Of course, if the group of students completing the questionnaires includes the most talented and hardworking students in the course, these would be the
Business Economics and Professional Accounting students to satisfy major requirements. Presumably, some of these students would not be very disappointed if, say, intermediate microeconomic theory were dropped from the list of major requirements. The 200 and 300 level major courses are taken by economics and accounting majors and students pursuing a concentration in finance, i.e. these courses are taken by volunteers who should know what they are in for. The average of the student rating for the overall effectiveness item is 2.54 (halfway between very good and good) for all 100 course sections. For 200 level required courses, 200 level major courses and 300 level major course, the overall rating is much lower (better); the ratings are 2.17, 1.9 and 1.81 respectively. The difference between the overall rating in the 100 level sections and the upper division sections is statistically significant. The median grade for all course sections taught in EFA is a 2.85 which is in the B/ B- range. The median grade for all course sections taught in the College is in the A-/B+ range. It is important to note that the median grade is based on the entire population of students who complete a course. This population is larger than the population completing the SEI forms. Again, the issue is whether the group of students completing the form is representative. Table 2 also provides data on the gender and the nationality of the course instructor. The percentage of courses taught by senior faculty members is also shown. In order to be promoted to the ranks of associate or full professor at our College a faculty member must demonstrate outstanding proficiency in the area of teaching. (Of course, there is also the possibility that once promoted the faculty member can rest on his or her laurels.) Finally, to account for the possibility that instructors with more experience are more effective, the table provides data on the number of semesters of prior teaching experience at Oneonta. In the next step of my research, I intend to collect data on the times of day at which courses are taught, the length of the class period, the rooms in which classes are taught and include a semester dummy variable. students who are best able to evaluate an instructor's effectiveness. Table Three also contains data on the level of the course. The

introductory 100 level courses have very heavy enrollments of students who come from outside the Division of Economics and Business and are taking the course to satisfy general education requirements or a related work requirement for their majors. These students might be more interested in an easy grade than in what they learn. The 200 and 300 level courses are taken mainly by Business Economics, Economics, and Professional Accounting majors. The 200 level required courses are the intermediate economic theory courses, corporate finance, and business law, which are taken by

Table 1: Summary Statistics Teacher Evaluation Questionnaire

Question	Mean	Standard deviation
Rigorousness of the course	2.27	.55
Organization of the Course	2.22	.67
Teaching skill is evidenced by classroom presentation	2.25	.76
How well instructional materials were coordinated with lecture	2.19	.69
Poise and self confidence	1.89	.59
Planning and clarity of examination questions, and their relevance to reading materials and classroom presentations.	2.29	.71
Ability to answer questions and clearly explain concepts	2.16	.73
Tolerance of attitudes and opinions of others	1.94	.61
Ability to maintain control of class discussion without getting into irrelevant matters	2.04	.61
Availability for consultation	1.99	.54
Maintains demanding grading practices	2.06	.45
Fairness in grading and in the implementation of his/her own announced grading procedure	2.12	.5
Overall evaluation of teaching effectiveness	2.26	.72

Table 2: Summary Statistics Class and Instructor Characteristics

Variable	Mean(Proportion)	Standard deviation
Enrollment	34.26	13.91
Withdrawals	2.53	4.55
Number of SEI forms completed	23.68	9.93
Percentage of survivors completing SEI form	75.98	15.49
100 level	39.35 (109/277)	
200 level required	34.66 (96/277)	
200 level major	6.5 (18/277)	
300 level major	19.49 (54/277)	
Median grade	2.85	.43
Male	68.23 (189/277)	
Non-native	21.66 (60/277)	
Rank (Assoc., full)	29.24 (81/277)	
Semesters of teaching experience at Oneonta	11.14	10.72

III. IS THE NOISE TO SIGNAL RATIO INFINITE?

To answer the question of the whether the data generated by the student evaluations are pure noise, I use two approaches. Of the nineteen faculty members who have taught courses in EFA over the last four years, eight have taught courses each semester. For each of these instructors, I have computed the weighted average (the weights are based on course enrollments) of the student responses to the "overall evaluation of teaching effectiveness" question. The results are shown in table 3.[2] The numbers in

parentheses are the rankings of the eight faculty members each semester (1= highest, 8= lowest).As a quick glance at the data in the table will confirm the SEI ratings are rather stable. For example, instructor E was the highest rated instructor each semester and instructor Q was the lowest rated. Each semester the difference between the SEI rating of the highest and lowest rated instructors is over two standard deviations, which is large. The correlations between the semester ratings are shown in table 4. The correlations between the semester ratings are generally too high to be the result of random chance. The correlations between the rankings tell the same essential story.

Table 3: SPI Scores by Semester (Ranking in parentheses)

Instructor	SP 12	F11	SP 11	F 10	SP 10	F 09	SP09	F08
A	1.8(4)	2.16(4)	1.75(3)	1.88(3)	2.03 (6)	1.89 (5)	1.83(4)	2.26 (5)
D	2.75(6)	2.79(6)	2.25(5)	2.41(6)	1.83(5)	2.21(6)	2.66(7)	2.09(4)
E	1.32(1)	1.4(1)	1.34(1)	1.22(1)	1.28(1)	1.3(1)	1.2(1)	1.31(1)
I	1.68(3)	1.51(2)	1.58(2)	2.03(4)	1.82(3)	1.79(3)	1.64(2)	1.98(3)
K	1.56(2)	1.79(3)	2.3(6)	1.42(2)	1.82(3)	1.63(2)	1.81(3)	2.43(6)
P	2.94(7)	3.1(7)	2.37(7)	3.31(7)	2.85(7)	2.61(7)	2.43(6)	2.64(7)
Q	3.03(8)	3.46(8)	3.07(8)	3.39(8)	3.33(8)	2.78(8)	2.69(8)	3.1(8)
Z	2.14(5)	2.31(5)	2.02(4)	2.13(5)	1.66(2)	1.88(4)	2.04(5)	1.86(2)

Table 4: Correlation between Semester SPI Scores

	SP 12	F11	SP 11	F 10	SP 10	F 09	SP 09	F 08
SP 12	1							
F 11	.9712	1						
SP 11	.803	.8647	1					
F 10	.9468	.9201	.7606	1				
SP 10	.7985	.8502	.8330	.9019	1			
F 09	.9602	.9577	.8353	.9824	.9241	1		
SP 09	.9555	.9411	.8574	.8518	.7307	.9103	1	
F 08	.6985	.7739	.8921	.7604	.9229	.8376	.7389	1

A different approach suggested by Watts and Bosshardt (cited in Isely and Sigh 2005) is to use a fixed effects model to capture instructor-related differences. The motivation is that is instructors may use different approaches to teach the same course. These differences could be difficult to measure. Instructors also differ on other hard to measure qualities such as charisma and sense of humor. Following Watts and Bosshardt, Isely and Singh and McPherson (2006), I create dummy variables for each instructor. (In this exercise, I exclude two faculty members who only taught at the college for one semester.) Using OLS, I regress the average of the student responses to the “overall evaluation of teaching effectiveness” on each on the instructor dummies. The results are presented in table 5. In general, the instructor coefficients are highly significant. For convenience, I provide the predicted instructor ratings in the last column of the table. Each instructor’s ranking is shown in parentheses.

Taken together, the two approaches indicate that the instructor ratings are not particularly noisy. The differences between the instructor ratings are large and tend to persist over time. Instructors who are highly rated tend to be highly rated consistently and similarly for instructors who receive low ratings.

The fact that our SEI instrument produces consistent rankings does not mean that it is any good or that it tells us what we want to know. The ratings could be affected by forces beyond the instructor's control such as the mix of courses taught, gender or nationality. They could also be influenced by forces under the instructor's control. For example, they could be influenced by an attempt to buy good student ratings with high grades or by an attempt to manipulate the student evaluation results by picking a propitious day on which to administer the evaluation questionnaire.

Table 5: OLS Regression Results with Instructor Dummy Variables

Variable	Estimated Coefficient	Standard Error	t-statistic	Estimated SPI Score (ranking)
Intercept	3.1177	.1189	26.2099**	
A	-1.2568	.1711	-7.3468**	1.86 (6)
C	-1.3821	.2279	-6.0657**	1.74 (4)
D	-.7904	.1536	-5.1454**	2.33 (10)
E	-1.8404	.1635	-11.2538**	1.28(1)
G	-.0178	.1864	-.0957	3.09 (15)
I	-1.4204	.1536	-9.2465**	1.69 (3)
J	-1.5256	.1683	-9.066**	1.59 (2)
K	-1.3157	.1549	-8.4913**	1.8 (5)
N	-.6915	.1864	-3.702**	2.43 (11)
P	-.3796	.1536	-2.4709*	2.74 (12)
Q	-.1735	.1549	-1.1199	2.94 (14)
U	-.3088	.1818	-1.6987	2.81(13)
V	-.8649	.1635	-5.2885**	2.26 (9)
Y	-.9795	.1742	-5.6231**	2.14 (7)
Z	-.9663	.1635	-5.7419**	2.14 (7)
AA	.0062	.2278	.0274	3.12 (16)

Number of observations=277
R^2 = .5926 adjusted R^2 = .5675 F = 23.635
** significant at the 99 percent confidence level
* significant at the 95 percent confidence level

IV. "THE STARS ARE STACKED AGAINST ME": THE IMPACT OF FACTORS BEYOND THE INSTRUCTOR'S CONTROL

A number of the variables included in table 2 can legitimately be viewed as beyond the instructor's control: course enrollment, the level of the course, gender, nationality and years of experience at the college. I use OLS to regress the "overall evaluation of teaching effectiveness" on these explanatory variables. The results are shown in table 6. While this set of variables has less explanatory power than the fixed effects model, the results show that the variables do matter. Course enrollment has a negative impact on the student ratings. A ten student increase in class size would cause a .17 point increase (deterioration) in the overall evaluation rating. The coefficients of the course level dummies for 200 level required courses and 300 level major courses show that students in the these upper division courses give

Table 6: OLS Regression Results for Course and Instructor Characteristics

Variable	Estimated coefficient	Standard Error	t-statistic
Constant	1.62	.2229	7.2696**
Enrollment	.0172	.0043	4.0396**
200 Required	-.3118	.1042	-2.9912**
200 Major	-.1946	.1888	-1.0309
300 major	-.4255	.1522	-2.7954**
Male	.3686	.0889	4.1437**
Foreign	.2541	.1177	2.1599*
Semesters at OSC	-.0048	.0039	-1.2275

Number of observations=277
R^2 = .3181 adjusted R^2 = .3003 F = 17.9234
** Significant at the 99 percent confidence level
+ Significant at the 90 percent confidence level

significantly lower (better) student ratings than do students in the 100 level introductory courses. This result confirms Dilts' (1980) finding. He found that student evaluations in required courses were .66 point worse than in courses taken by majors. The average rating for a male instructor was .37 point higher (worse) than that of a female instructor. The average rating of a foreign born instructor was .25 point higher than the rating for a native-born instructor. These differences are statistically significant. Interestingly, instructor experience as measured by semesters at the college has a negligible and statistically insignificant on the student ratings.

Of course, these results are only suggestive. For example, female instructors might receive higher ratings than their male colleagues not because of their gender but because they are more effective teachers. It is also possible that female instructors receive higher ratings because they are more lenient graders than their male colleagues. Therefore, it is important to consider the impact of factors under a faculty member's control on the SEI results.

V. "THE FAULT LIES NOT IN THE STARS BUT IN OURSELVES": THE IMPACT OF FACTORS UNDER THE INSTRUCTOR'S CONTROL.

In any consideration of factors under the faculty member's control, the obvious place to start is with grades. If the "overall effectiveness" ratings are regressed on the median grade in each course, the coefficient of the median grade is -.59, which means that a 1 point increase in the median GPA in a course section lowers (improves) the "overall effectiveness "by .59 point. The coefficient is statistically significant. This result is suspect on a number of grounds. I'll consider one in particular. In the 100 level courses, the median grade is 2.56 (B-/C+) which is significantly lower than the median grades in 200 level required courses (2.9, B/B-), 200 level major courses (3.19, B/B+), and 300 level major courses (3.21, B/B+). These differences are easy to account for. The 100 level courses attract heavy enrollments of students from outside the Division of Economics and Business who take the course to satisfy related work requirements for their majors or general education requirements. Unlike students in the upper division classes who must earn a C or better in order to avoid repeating the class, students from outside the division taking the 100 classes only need to earn a D- to receive credit for the course. To adjust for the impact of course level on grades, I regress the overall evaluation rating on course level and median grade. These results are shown in table 7.

As is readily apparent, once we control for the level of the course, the median grade has a statistically insignificant impact on the student rating.

Table 7: OLS Regression Results for Course Level and Median Grade

Variable	Estimated coefficient	Standard Error	t-statistic
Intercept	3.0397	.2918	10.4173**
200 Required	-.4383	.0974	-4.5007**
200 Major	-.6397	.1761	-3.633**
300 major	-.8234	.1275	-6.4596**
Median Grade	-.1484	.1114	-1.3321

Number of observations=277
R^2 = .2457 adjusted R^2 = .2346 F = 22.1473
** Significant at the 99 percent confidence level
+ Significant at the 90 percent confidence level

VI. CONCLUDING COMMENTS

At this stage, I can offer some preliminary answers to my three questions. First, the results of the student evaluations are not noisy. The results are stable from semester to semester. Second, the results do appear to be influenced by factors beyond the instructor's control such as the level of the course. The results do not seem to be influenced by grades. At least in my department, a policy of assigning easy grades does not appear to be winning approach to receiving high scores on the student evaluations. The results appear to be driven by what goes on in the classroom. Exactly what differentiates more effective from less effective will be explored in the next phase of the study.

ENDNOTES

1. A more detailed review of the literature can be found in O'Dea (2010).
2. In assigning letters to identify faculty members, I did not use any letters that correspond to the first initial of a faculty member's last name. For example, no faculty member is identified by the latter O. The remaining letters were assigned to faculty members randomly.

REFERENCES

Becker, William E. 1997. "Teaching Economics to Undergraduates." *Journal of Economic Literature* , 35(3): 1347-1373.

Becker, William E., and Michael Watts. 1999. "How Departments of Economics Evaluate Teaching." *The American Economic Review,* 89(2): 344-349.

Boex, L. F. Jameson. 2000. "Attributes of Effective Economics Instructors: An Analysis of Student Evaluation." *The Journal of Economic Education,* 31(3): 211-227.

Bosshardt, William, and Michael Watts. 2001. "Comparing Student and Instructor Evaluations of Teaching." *The Journal of Economic Education,* 32(1): 3-17.

DeCanio, Stephen J. 1986. "Student Evaluation of Teaching: A Multinomial Logit Approach." *The Journal of Economic Education,* 17(3): 165-176.

Dilts, David A. 1980. "A Statistical Interpretation of Student Evaluation Feedback." *The Journal of Economic Education,* 11(2):10-15.

Isely, Paul and Harinder Singh. 2005. "Do Higher Grades Lead to Favorable Student Evaluations?" *The Journal of Economic Education,* 36(1): 29-42.

Marlin Jr., James W. and James F. Niss. 1980. "End of Course Evaluations as Indications of Student Learning and Institutional Effectiveness." *The Journal of Economic Education, 1*1(2): 16-27.

McPherson, Michael A. 2006. "Determinants of How Students Evaluate Teachers." *The Journal of Economic Education,* 37(1): 3-20.

Nelson, Jon P., and Kathleen A. Lynch. "Grade Inflation, Real Income, Simultaneity, and Teaching Evaluations." *The Journal of Economic Education,* 15(1): 21-37.

O'Dea, William P. 2010. "A Chair's Guide to Student Evaluations." *Proceedings of the New York State Economics Association*, 3: 115-125.

Siegfried,John J., and Rendig Fels. 1979. "Teaching College Economics: A Survey." *Journal of Economic Literature*, 17(3): 923-969.

Weinberg, Bruce , Masanori Hashimoto, and Belton M. Fleisher. 2009. "Evaluating Teaching in Higher Education." *The Journal of Economic Education, 40*(3): 227-261.