Student Evaluation Of Teacher Performance: Random Pre-Destination

Kim L. Chuah, (E-mail: chuahki@andrews.edu), Andrews University
Cynthia Hill, (E-mail: hillcynt@isu.edu), Idaho State University

Abstract

The student evaluation, used to measure students' perceptions of teacher performance, has been increasingly used as the predominant component in assessing teaching effectiveness (Waters et al. 1988), and the widespread movement of outcomes assessment across the country makes this trend likely to continue in the future (McCoy et al. 1994, AACSB 1994, SACS 1995). Substantial research has been conducted with regard to the reliability and accuracy of student evaluation of teaching quality, and a considerable number of uncontrollable factors are found to bias the results of the evaluation rating. This paper identifies one more factor. Each student has an "evaluator profile", which decreases the reliability of the student evaluation. An "evaluator profile" is a persistent pattern of evaluating behavior that may or may not be consistent with the quality of the characteristic being evaluated. Each class of students consists of a random sample of different evaluator profiles. A student evaluation rating of a teacher's performance is biased up or down depending on the concentration of high or low evaluator profiles present. This paper further shows through simulation the degree to which student "evaluator profiles" impact the overall student evaluation rating of teacher performance. We find that there is evidence to support the "evaluator profile" conjecture, and that these "evaluator profiles" do in fact have the potential to change overall student evaluation ratings substantially.

Introduction

An overwhelming majority of universities and colleges across the nation have students evaluate their instructor's teaching effectiveness at the end of each semester (Seldin 1989). This student evaluation, used to measure students' perceptions of teacher performance, has been increasingly used as the predominant component in assessing teaching effectiveness (Waters et al. 1988), and the widespread movement of assessment across the country makes this trend likely to continue in the future (McCoy et al. 1994, AACSB 1994, SACS 1995). This simple and convenient administrative tool for teacher quality assessment often enters into the decision making process for faculty tenure, promotion, and merit pay (White 1995, Dilts and Faterni 1982, Costin et al. 1971). Therefore the importance of an accurate and reliable student evaluation is compelling.

Although the ease with which the student evaluation can be administered certainly adds to its continued dominance in assessing teaching effectiveness, reliability and accuracy are not strong points. The research that has been conducted with regard to the reliability and accuracy of student evaluation of teaching quality is considerable. Numerous factors have been found that bias the results of the evaluation rating. Some research suggests that instructor characteristics such as: gender, grading leniency, teaching experience, and mode of teaching may skew student evaluation ratings (Husbands 1996, Fandt and Stevens 1993, Crader and Butler 1996, and Greenwald and Gilmore 1997). Class characteristics have also been found to bias evaluations; these include: class size, course level, requirement or elective status of the course, time-of-day of class meetings, and academic discipline (Danielsen and White 1976, Mateo and Fernandez 1996, Mirus 1973). Individual student characteristics such as student expectations regarding instructors, student gender, student rank in school, prior subject interest, expected grade, emotional state, student age, and student ability have all been shown to color student evaluation ratings (Kulikand and McDeachie 1975, Kelly 1972, Nichols and Soper 1972, Seiver 1982, Nelson and Lynch 1984, Seiver 1982). Finally, Abrami et al. (1990) and Gramlich and Greenlee (1993) find very little positive correlation between high

student evaluations and high comprehensive exam scores of these same students. Further, factors that are out of the instructor's control have considerable impact on the evaluation rating. Thus, there is considerable reason for concern with regard to the reliability and accuracy of the student evaluation.

In addition to the factors previously identified, this paper identifies one more factor, a student's "evaluator profile", which decreases the reliability of the student evaluation of teaching quality. It is the hypothesis of this paper that each student has an "evaluator profile", or a consistent pattern of evaluation, which may or may not be consistent with the quality of the teaching characteristics being evaluated. An individual student's "evaluator profile" may be consistently low compared to the mean for the class, or it may be consistently high compared to the mean, or it may be directly correlated with the mean (increasing when the mean increases, and decreasing when the mean decreases). A student's "evaluator profile" contributes to the overall bias of the student evaluation rating in that the "evaluator profile" is relatively stable over time, and is therefore not necessarily directly correlated with the quality of instruction.

The implication can be disturbing. The composition of the particular class (i.e., the number of "low evaluators" or "high evaluators") biases the student evaluation rating. An instructor's evaluation rating will depend upon the number of students in that particular class with particular types of evaluator profiles. This paper also shows, through simulation, the degree to which students' "evaluator profiles" may impact the overall student evaluation rating of teaching effectiveness. For example, if an instructor has a large number of "low evaluators" for one particular class, then his or her overall student evaluation rating will be lower than "normal" simply due to the composition of the class, not due to the quality of instruction. On the other hand, if an instructor has a large number of "high evaluators" for a particular class, then his or her evaluation rating will be higher than "normal", once again based on the composition of the class, not based on the quality of instruction.

Methodology and Data

In order to examine the hypothesis of a student's "evaluator profile" as described above, information was needed from a random group of students regarding their relative evaluations over time and over a number of instructors. In other words, we needed to follow a group of students over several semesters and examine the ways in which each individual student evaluated instructor effectiveness. The anonymity associated with student evaluations made obtaining this type of data nearly impossible. This task was further complicated due to the necessity of matching individual student evaluations in a number of classes over a number of semesters. Due to these seemingly insurmountable difficulties, a comparable data set was therefore developed instead.

This data set was drawn from student-evaluations of peer presentations. The presentations were performed in business and economics classes where students evaluated one another. In each of four different business and economics classes, students were required to form groups of four to nine students, prepare a research project, and present their information to the class in some type of formal presentation at the end of the semester. The students were then required to evaluate one another, and the instructors used the peer evaluations to determine the majority of the project/presentation grade. Hence, this alternative data set provided two important perspectives: how a student evaluates a series of presentations, and how the distribution of these evaluations affects the overall rating of an individual presentation.

Class sizes from which the data were drawn ranged from 33-70. Two of the four classes were offered at a public university whereas the other two classes were offered at a private university. Two of the classes were Principles of Macroeconomics, one was a lower division International Business Environment course and the other was an upper division Productions and Operations Management course. The group project and presentation was a requirement for all of the above listed classes and worth approximately 15% of the final grade. The students were informed throughout the semester that their peers would evaluate their presentations, and that the majority of the project/presentation grade would be dependent upon the average rating received on their presentation.

The presentations were graded on a scale of 1(lowest) - 5 (highest) based on the five characteristics listed below. All four classes received identical evaluation forms. Students were asked to evaluate the presentations using the following criteria:

Content:	Depth, substance, and completeness of the material. How much work did the group put into the project as revealed through the content and accuracy of the presentation?
Professionalism:	Well organized, and well prepared, free from major grammatical, spelling, and typographical errors. How much attention did this group pay to being detailed and professional?
Clarity:	Ease of legibility, ease of interpretation, coherence, and flow. Did this group get their main points across in a clear and straightforward manner?
Style:	Originality, creativity, flair, and enthusiasm. Did this group have interest in the subject? Was their presentation format original or interesting?
Knowledge Gained:	Furnished additional information compared to the information presented during regular class time. How much knowledge was gained from this presentation over and above what had already been learned throughout the semester?

These five characteristics were developed to imitate the integral components of a student evaluation of teaching effectiveness. Course content, relevance and usefulness of course content (Content), instructor's preparation, course organization (Professionalism), clarity of course objectives, clarity of student responsibilities and requirements (Clarity), instructor's use of examples and illustrations, instructor's enthusiasm, interest level of class sessions (Style), instructor's contribution to the course, use of class time, and amount learned in the course (Knowledge Gained) are all key objectives which instructor evaluation instruments attempt to measure. The five characteristics listed are an attempt to mimic these evaluation objectives in a condensed framework.

The data set was coded in the following manner. Each student was provided with a numerical ID number to preserve anonymity. The average score of the above five characteristics, denoted r_{ij}, was the evaluation by student i on the presentation of group j. An "individual rating average", r_i, was computed for each student evaluator as follows:

$r_i = (\Sigma_j r_{ij})/J$, where J = number of presentations evaluated by student i.

This "individual rating average" provided a possible glimpse into the evaluator profile of the individual student. The "individual rating average" was then arranged from lowest to highest, and divided into four ranked groups: extremely low evaluators, low evaluators, high evaluators, and extremely high evaluators.

For each group presentation, the list of student evaluations, r_{ij}'s were arranged from lowest to highest and four similarly ranked groups were identified. All equivalently ranked groups were compared to find individuals common to these groups. For example, the extremely low evaluators for the various presentations were compared to see if there existed a student or a number of students that belonged to the extremely low evaluators group for a two-thirds of the presentations. Similar comparisons were made for the other groups.

Based on the distribution exhibited by the average scores of students in ranked groups, fourteen samples of random numbers were generated to simulate the various possible outcomes of a class of thirty-five students. The average score of each sample indicated an evaluation outcome for a class of thirty-five students. The fourteen samples showed how much the average score could deviate due to the random occurrences of "extremely low evaluators" or "extremely high evaluators" in a particular class.

Analysis of the Data

Table 1 shows the results of compiling the comparisons of individuals in the specific evaluator categories. The boundaries separating the categories in each presentation was calculated using the mean score of the presentation and the standard deviation. The mean separated evaluators into two unequal groups, the low and extremely low evaluators in one group, and the high and extremely high evaluators in the other group. The former group was further divided into the low evaluators group and the extremely low evaluators group using the boundary obtained by taking the mean minus one standard deviation. Similarly, the latter group was further divided into the high evaluators and extremely high evaluators group by calculating the boundary as the mean plus one standard deviation.

Table 1: Evaluator Profile Distributions

Category	School A	School B
	Percent in category	
Extremely low evaluators	19.2%	16.1%
Low evaluators	17.9%	8.9%
Extremely high & high evaluators	53.3%	47.5%

Note that each presentation could have different boundaries for these ranked groups according to its mean score and standard deviation. Student evaluators could be in different ranked groups in different presentations. It was determined that of those students who belonged to the extremely low evaluators group in at least one presentation, 19.2% of those students consistently fell in this category in School A, and 16.1% consistently fell in this category in School B. A student was considered an extremely low evaluator if the student fell into that category for two-thirds of the presentations. Using this same criteria, 17.9% of students in School A and 8.9% in School B consistently fell in the in the low evaluators category. Because students evaluated on average higher than the mid point of the scale, the boundary separating "high evaluator" and "extremely high evaluator" often exceeded the maximum of the scale. This phenomenon "squeezed" out the extremely high evaluators group in several of the presentations. Further, a number of the students consistently provided an evaluation score of 5.0, a perfect score, in most of their evaluations. This indicates that they were the extremely high evaluators that this paper seeks to identify. In order to capture their existence, the extremely high evaluators and high evaluators groups were combined. Of those students who belonged to this combined group, 53.3% of them in School A and 47.5% of them in School B consistently fell into this category. The rest of the students belonged in two or more categories, but never more than two-thirds of the time in any one category.

Simulation Results

Using the distribution exhibited by the "rating averages", 35 random numbers were generated to simulate a sample of individual students belonging to the four ranked groups. Each simulated student was then assigned the mean rating average that was representative of their group as their own rating average. These constituted a simulated sample. The average of the sample represented the evaluation rating received by an instructor from a class. The process was repeated 14 times to simulate the numerous configurations of distribution of the evaluator profile. Specifically, 14 samples were generated for each of the schools. The results are summarized in Table 2.

The results from both schools were strikingly similar, although not identical. The distributions from the two schools were remarkably close. According to the results in table 2, if such groups of consistent evaluators existed and if the distribution existed as exhibited, the possible bias could result in mean rating average as high as 4.43, or as low as 4.00 in School A, and between 4.30 and 3.96 in School B. This showed a possible difference of 0.43 and 0.34 respectively, depending on which sample of students actually occurred.

Table 2: Results of Random Sample Simulations Using Profile Distributions

	School A				School B			
Rating average	4.2811				4.1903			
Std.deviation	0.5658				0.4390			
	boundaries		number	%	boundaries		number	%
Extremely low	1.0000	3.7154	10	18%	1.0000	3.7514	11	17%
Low	3.7154	4.2811	15	26%	3.7514	4.1903	19	30%
High	4.2811	4.8469	21	37%	4.1903	4.6293	22	34%
Extremely high	4.8469	5.0000	11	19%	4.6293	5.0000	12	19%
Total			57				64	
Simulation Sample	Simulated Mean				Simulated Mean			
1	*4.33*				*4.23*			
2	*4.03*				*4.00*			
3	*4.27*				*4.18*			
4	*4.24*				*4.16*			
5	*4.39*				*4.27*			
6	*4.20*				*4.12*			
7	*4.31*				*4.20*			
8	*4.06*				*4.02*			
9	*4.39*				*4.30*			
10	*4.27*				*4.19*			
11	*4.10*				*4.02*			
12	*4.27*				*4.16*			
13	*4.43*				*4.28*			
14	*4.02*				*3.96*			
Simulated mean =	*4.2351*				*4.1488*			
Simulated standard error =	*0.1367*				*0.1099*			
High =	*4.43*				*4.30*			
Low =	*4.02*				*3.96*			
Range =	*0.41*				*0.34*			

Conclusion

Using two sets of data from two different universities, we have shown that the existence of stable evaluator profiles can bias a teacher's performance evaluation. This stable profile is not consistent with evaluating the true performance of the teacher. The bias may add as large a difference of 0.43, which can be in either direction. In the "long run" the randomness of the distribution of these evaluators may be averaged out, as in our simulation of fourteen samples. Hence a general picture of performance may emerge over time. But at any point in time only one sample exists, and the evaluation ratings may be biased downward.

Our conclusions add strength to the caution raised by other researchers. The student-generated evaluations on teacher performance are not reliable because there are numerous uncontrollable factors that are not correlated with quality of teaching. Even when ratings are collected over as many as three to five years, the time frame may not be sufficient to remove the bias. When a school is small and students enrolled in the numerous classes taught by a specific teacher are coming from a small pool of students in the same discipline, the lack of randomness becomes an issue. In other words, similar pockets of low evaluators may enroll in the classes of an instructor repeatedly over the years. This will exaggerate the problem.

This paper only shows the existence of such evaluators, and the degree of potential impact. More research

may be needed to ascertain the degree and severity associated with this evaluator bias and how this evaluator bias can be addressed. As previously mentioned, student privacy issues prevent a more direct study of the problem.

References

1. AACSB (American Assembly of Collegiate Schools of Business -- The International Association for Management Education). Achieving Quality and Continuous Improvement Through Self-Evaluation and Peer Review. *Standards for Business and Accounting Accreditation*. St. Louis, Missouri. 1994.
2. Abrami, P.C., S. d'Apollonia, and P.A. Cohen, "Validity of Student Ratings of Instruction: What We Know and What We Do Not." *Journal of Educational Psychology*, June 1990, pp. 219-231.
3. Costin, Frank, William T. Greenough, and Robert J. Menges. "Student Ratings of College Teaching: Reliability, Validity, and Usefulness," *Review of Educational Research*, December 1971, pp. 511-535.
4. Danielsen, A.I., and White, R.A. "Some Evidence on the Variables Associated with Student Evaluations of Teachers," *Journal of Economic Education*, Spring 1976, pp. 117-119.
5. Dilts, David A. "A Statistical Interpretation of Student Evaluation Feedback," *Journal of Economic Education*, Spring 1980, pp. 10-15.
6. Gramlich, Edward M., and Glen A. Greenlee. "Measuring Teacher Performance," *Journal of Economic Education*, Winter 1993, pp. 3-13.
7. Kelley, Allen C. "Uses and Abuses of Course Evaluations as Measure of Educational Output," *Journal of Economic Education*, Fall 1982, pp. 13-18.
8. Kulik, J.A., and McKeachie, W.J. "The Evaluation of Teachers in Higher Education," In E.N. Kerlinger (Ed.), *Review of Research in Education*, 1975, 3, pp. 210-240. Itasca, NY: Peacock.
9. Mateo, M.A., and Fernandez, J. "Incidence of Class Size on the Evaluation of University Teaching Quality," *Educational and Psychological Measurement*, 1966, 56, pp. 771-778.
10. McCoy, James P., Don Chamberlain, and Rob Seay. "The status and preception of University Outcomes Assessment in economics," *Journal of Economic Education*, Fall 1994, pp. 358-366.
11. Mirus, Rolf. "Some Implications of Student Evaluations of Teachers." *Journal of Economic Education*, Fall 1973, pp. 35-37.
12. Nelson, Jon P., and Kathleen Lynch. "Grade Inflation, Real Income, Simultaneity, and Teaching Evaluations," *Journal of Economic Education*, Winter 1984, pp. 21-37.
13. Nichols, Alan, and John C. Soper. "Economic Man in the Classroom," *Journal of Political Economy*, September/October 1972, pp. 1069-1073.
14. SACS (Southern Association of Colleges and Schools). *Criteria for Accreditation: Commission on Colleges*. Decatur, Georgia. 1995.
15. Seiver, Daniel A. "Evaluations and Grades: A Simultaneous Framework," *Journal of Economic Education*, Summer 1982, pp. 32-38.
16. Seldin, P. "How Colleges Evaluate Professors." *American Association for Higher Education Bulletin* 41, March 1989, pp. 3-7.
17. Waters, M., Kemp, E., & Pucci, A. "High and Low Faculty Evaluations: Descriptions by Students," *Teaching of Psychology*, 1988, 15, 203-204.
18. White, Lawrence J. Efforts by Departments of Economics to Assess Teaching Effectiveness: Results of an Informal Survey," *Journal of Economic Education*, Winter 1995, pp. 81-85.