Estimating Determinants of Student Evaluation Scores to Improve Teaching

Laura McCann[1] and Michael Burton[2]

Contact addresses:

[1] Department of Agricultural Economics
University of Missouri
200 Mumford Hall
Columbia, MO 65211-6200
U.S.A.
McCannL@missouri.edu

[2]Agricultural and Resource Economics
University of Western Australia
35 Stirling Highway
Crawley, WA 6009
AUSTRALIA
Mpburton@agric.uwa.edu.au

Selected Paper for the AAEA Annual Meeting, Chicago, August 5-8, 2001.

Keywords: Resident Instruction, Student Evaluations

Abstract

Student evaluations are used for both formative and summative assessment of teachers. This paper provides a method to make more effective use of these student evaluations by individual teachers. Data on three years of evaluations in two courses were used to develop regression models to explain overall effectiveness of teaching. The relative importance of explanatory variables changed with the course taught.

Introduction

Student ratings can be used for both formative and summative assessment of teachers in the same way that exams are used both to provide feedback to students so they can improve, and to evaluate their performance. In a meta analysis, Cohen (1981) found that the correlation between course ratings and mean student achievement was 0.47. He also found that global ratings, such as overall effectiveness, correlated more highly with student learning than more specific items. Therefore, while student evaluations do not provide a perfect measure of student learning, they are indicative. In any case, promotion and tenure committees generally rely on student evaluations rather than objective measures of learning. This paper provides a method to make more effective use of these student evaluations in order to: 1) improve one's teaching and subsequent evaluations, 2) provide a framework for reflection which can be incorporated into a teaching portfolio.

Student evaluations of teaching have been used for a number of purposes (Ory 2000). In the 1960s, they were promoted by students to improve public accountability and to help students made decisions regarding courses. In the 1970s, the primary focus was on development of faculty. In the 1980s and 1990s, teaching evaluation has been extensively used by administrators for promotion and tenure decisions. In a study of

deans of liberal arts colleges, Seldin (1999a) reported that 55% of deans in 1978 said systematic student ratings were always used in evaluating teaching performance, in 1998 this number was 88%. This is due to an increasingly litigious climate on campuses which requires that decisions are made on objective criteria (Ory 2000, Seldin 1999b). Fortunately, hundreds of studies have determined that, overall, student ratings are both reliable and valid (Seldin 1997). A more recent trend is the use of a wider variety of measures of teaching effectiveness in making personnel decisions and a more structured and systematized process for collecting data (Seldin 1999a). For example, self-evaluation was always used by 37% of deans in 1978, and by 59% in 1998, while the use of course syllabi and exams increased from 14% to 39% over this period.

Teaching portfolios are increasingly being used to evaluate teaching performance. Zubizarreta (1999 p. 164) defines a teaching portfolio as "an evidence-based written document in which a faculty member strategically organizes concise, selective details of current teaching accomplishment and uses such information for documentation of performance but more significantly for reflective analysis and peer collaboration leading to improvement of teaching and student learning". Reflective analysis of student ratings and a description of how it was used to improve teaching can be one component of a teaching portfolio.

Seldin (1999a) indicated that there are more than 15,000 studies on teaching effectiveness. There is consensus on the characteristics of good teaching. In a survey article, Eble (1988 p. 21) found that "Most studies stress knowledge and organization of subject matter, skills in instruction, and personal qualities and attitudes useful to working

with students." A less dry definition was put forth by Miller (cited in Seldin 1999b): "Effective teachers personify enthusiasm for their students, the area of competence, and life itself. They know their subject, can explain it clearly, and are willing to do so." On the more limited topic of student evaluation of teaching performance, there are 2175 references (Cashin 1999). McKeachie (1975) found that ratings can lead to improvement in teaching if 1) ratings revealed something new to the teacher, 2) the teacher was motivated to improve, and 3) the teacher knew how to improve. Open ended questions as well as diagnostic questions or questions on specific teaching behaviors should be included in student evaluation forms to increase their usefulness for faculty development (Seldin 1997, Cashin 1999). Comments on evaluation forms provide detailed input as to specific ways in which teaching can be improved.

An EconLit search using the key words "student evaluations or student ratings" found 21 articles dating back to 1972. Most of these papers focus on the determinants of the rating for a global item such as teaching effectiveness. They combine a number of instructors and courses so one is left with knowing that, in general, certain variables, such as organization and clarity, affect teaching effectiveness (Boex 2000). The focus of this paper however is to demonstrate that, instead of relying on the literature, teachers need to analyze their own student ratings in order to increase their effectiveness as a formative assessment tool. Teachers need to be able to answer the question: "What should I do to improve?" in addition to "What characteristics are associated with good teachers?". The average scores for items on student evaluations don't provide any indication as to which areas are more important to the overall level of student satisfaction with the course and thus provide little guidance as to how to allocate one's teaching improvement effort.

Data

Data on three years of evaluations in two courses were used to develop regression models to explain overall effectiveness of teaching as a function of a number of other variables including year taught, clarity of explanations *(Explanations)*, organization of the class sessions (*Organization)*, extent to which the students have used learning opportunities *(Opportunity)*, the teachers knowledge of the subject matter *(Knowledge)*, enthusiasm for teaching the subject *(Enthusiasm)*, concern for students (*Concern)*, and delivery pace *(Pace)*. Responses were on a Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). In each case, the students were non-economics majors and the courses were required, both of which tend to reduce student ratings (McKeachie 1986).

Only the first four weeks of Environmental Policy (EP) relate to environmental economics so this is the period covered by the evaluations. It is taught in first semester. This course usually has an enrollment of approximately 80 students who come from a variety of disciplines, primarily Environmental Engineering, Environmental Science, Geography, and Natural Resource Management. Some students had taken or were concurrently taking a course called Environmental Economics. From conversations with students, it became obvious that some of them had never had an economics course in either high school or university while others had been exposed to supply and demand curves which form the basis for the pollution abatement diagrams used in environmental economics texts.

Economics for Agriculture and Resource Management (EARM) is taught by the first author for the entire second semester. There are approximately 50 students in the course each year essentially all of whom are majoring in Agriculture, Horticulture or Natural

Resource Management. Students had all taken an applied macroeconomics course the previous year. Therefore, compared to EP, the students in EARM are much more homogeneous both as far as their interests and economics backgrounds.

Given that EARM is taught after EP, a number of improvements were incorporated into EARM the first time it was taught, based on the comments from EP. For example, comments from the open-ended questions indicated that the students wanted to see more Australian examples so an effort was made to do this for EARM. In addition a "one minute paper" (Angelo and Cross 1993) was incorporated into EARM in which students are asked to write down a question they have or mention something that they have learned in the course. This serves the purpose of providing early, informal assessment so that improvements can be made for the current students. Another modification made in response to EP comments was to include a *pace* item in the evaluation form for EARM and to present graphs at a slower pace.

A number of changes were also made to EP the following year in response to the EP evaluations. More Australian examples were used, fewer types of graphs were presented, graphs were explained more slowly, less detail was presented, a special tutorial session for students who hadn't taken economics previously was conducted, an overview of the importance of environmental economics in the context of environmental policy was included at the beginning, a handout on graphs was provided, political economy was included to reduce overlap between EP and Environmental Economics, an Australian book on environmental economics designed for people without an economics background was included as a text, as well as a number of improvements related to explanations of specific concepts. Due to these targeted changes, the rating for *Explanations* in EP was

significantly ($p < 0.004$) higher in 1999 and 2000 than in 1998 resulting in significantly higher ratings for *Effectiveness*.

Results and Discussion

While ordinary least squares (OLS) exhibits a number of problems with respect to a categorical dependent variable, (Judge et al. 1985, DeCanio 1986) it has the advantage of ease of interpretation. A preliminary OLS regression analysis found that, as with previous studies (McKeachie 1986), gender was not a significant explanatory variable. Since not all students indicated their gender, thus lowering the number of observations, this variable was deleted for subsequent analyses. Correlations between the explanatory variables were all below 0.50 so multi-collinearity was not a problem as it is for studies where there are a much larger number of related explanatory variables (Boex 2000). For Environmental Policy (n = 156), clarity of *Explanations* and *Organization* of the course sessions were statistically significant at the 0.05 level (Table 1). The coefficient for *Explanations* (0.59) was much larger than that for *Organization* (0.17) indicating that more effort should be spent on improving explanations than improving organization. Year the course was taught, extent to which the students have used learning *Opportunities*, the teachers *Knowledge* of the subject matter, *Enthusiasm*, and *Concern* for students were not significant. For EP, *Explanations* was the explanatory variable with the lowest average score.

For EARM (n=111), *Explanations, Enthusiasm*, and *Pace* were all significant at the 0.05 level, while *Knowledge* was significant at the 0.10 level (Table 2). *Enthusiasm* had the largest coefficient (0.41), followed by *Explanations* (0.31), *Pace* (0.22) and *Knowledge* (0.16). It is important to point out that while these coefficients may be used to efficiently

allocate effort, they should be used in conjunction with the average evaluation score since in this case, the score for enthusiasm was the highest of all the variables so there was little room for improvement.

Table 1. Explanation of Teaching Effectiveness in Environmental Policy, OLS

	Coefficient	Standard Error	t-Statistic	P-value
Intercept	-0.39	0.39	-1.00	0.32
1999 (base is 1998)	0.04	0.13	0.27	0.78
2000	0.04	0.13	0.31	0.76
Explanations	0.59	0.06	10.51	0.00
Organization	0.17	0.07	2.57	0.01
Opportunity	0.08	0.06	1.43	0.15
Knowledge	0.03	0.07	0.44	0.66
Enthusiasm	0.13	0.09	1.49	0.14
Concern	0.03	0.08	0.42	0.68

Adjusted $R^2 = 0.62$

Table 2. Explanation of Teaching Effectiveness in Economics for Agriculture and Natural Resource Management, OLS

	Coefficient	Standard Error	t-Statistic	P-value
Intercept	-0.70	0.42	-1.69	0.09
1999 (base is 1998)	-0.01	0.14	-0.04	0.97
2000	0.03	0.14	0.22	0.83
Explanations	0.31	0.08	3.80	0.00
Organization	0.07	0.08	0.93	0.35
Opportunity	-0.05	0.06	-0.84	0.40
Knowledge	0.16	0.09	1.84	0.07
Enthusiasm	0.41	0.10	4.17	0.00
Concern	0.00	0.08	0.05	0.96
Pace	0.22	0.09	2.42	0.02

Adjusted $R^2 = 0.62$

A pooled data set was created for EP and EARM[1]. Since year was not significant in either model, it was deleted from subsequent analysis. Since *Pace* was not available for EP, this was also deleted from the pooled data set. Preliminary analysis showed that the

[1] All estimation and interpretation has used Stata Version 6 (StataCorp, 1999).

coefficients were not significantly different for the two courses except for the variables *Explanations* and *Enthusiasm.* Therefore the final model estimates them separately. *Explanations* was significant at the 0.05 level for both courses but the coefficient was higher for EP. *Enthusiasm* was also significant for both courses although only at the 0.10 level for EP and the coefficient was larger for EARM. Other authors have found that determinants of teacher effectiveness vary across courses (Abrami 1989, Boex 2000). *Organization* was also significant at the 0.05 level while *Knowledge* was significant at the 0.10 level. *Concern* and *Opportunity* weren't significant. The adjusted R^2 for all the models was over 0.60 which is quite high given the limited number of variables that were included and the lack of student characteristics in the model. DeCanio (1986) had 5615 observations and 21 explanatory variables and obtained an R^2 of 0.73 for an OLS model. This would indicate that even for an individual, a few courses over a few years can provide sufficient information to improve one's teaching.

Table 3. Explanation of Teaching Effectiveness for the Pooled Data Sets, OLS

	Coefficient	Standard Error	t-Statistic	P-value
Intercept	-0.29	0.26	-1.12	0.26
Explanations EP	0.60	0.05	12.05	0.00
Explanations EARM	0.39	0.07	5.28	0.00
Organization	0.13	0.05	2.63	0.01
Opportunity	0.03	0.04	0.75	0.45
Knowledge	0.10	0.05	1.81	0.07
Enthusiasm EP	0.13	0.07	1.95	0.05
Enthusiasm EARM	0.38	0.08	4.70	0.00
Concern	0.02	0.05	0.38	0.70

Adjusted $R^2 = 0.64$

The data being considered is strictly ordinal, therefore an ordinal logit model was also employed. A linear relationship is assumed between the explanatory variables and an

unobserved, latent, variable y^*:

$$y^* = \beta' x + \epsilon$$

where β' is a vector of parameters, x the explanatory variables and ϵ the error term. Since there are 5 categories, it is further assumed that the observed data is generated by the process:

$y = 1$ if $y^* \leq \mu_1$
$y = 2$ if $\mu_1 < y^* \leq \mu_2$
$y = 3$ if $\mu_2 < y^* \leq \mu_3$
$y = 4$ if $\mu_3 < y^* \leq \mu_4$
$y = 5$ if $\mu_4 < y^*$

where μ_i are a set of cut points that have to be estimated. A number of nested models were developed, and only the final model is reported here (Table 4). Of particular interest are any differences between courses in the relationship between specific evaluation items and the measure of overall effectiveness. Restricting the parameters and cut-points to be the same for both courses was rejected when compared to estimating separate models for each (test statistic of 25.41 compared to $\chi^2_{10,0.05} = 18.31$). However, differences between the two courses seems to be restricted to the *Explanation* and *Enthusiasm* variables. An alternative restricted model, which imposes common cut points and common parameters for all other variables is accepted, as compared with the general model (test statistic of 9.63 compared to $\chi^2_{8,0.05} = 15.51$). These results are reported below. It should be noted that neither the gender of the student nor the year in which the survey was conducted was significant in any specification, similar to the OLS models. Note that the latter does not imply that there has been no change in the level of

effectiveness over time, but that there is no change in the relationship between individual measures of performance and the measure of overall effectiveness.

Table 4 Ordered Logit Estimates for Pooled Data Sets. (n=267)

Factor	Coefficient	Standard Error	z	P-value
Explanation (EP)	1.89	0.21	9.08	0.00
Explanation (EARM)	1.22	0.28	4.44	0.00
Organization	0.46	0.18	2.53	0.01
Opportunity	0.21	0.16	1.32	0.19
Knowledge	0.32	0.21	1.52	0.13
Enthusiasm (EP)	0.62	0.25	2.43	0.01
Enthusiasm (EARM)	1.41	0.31	4.49	0.00
Concern	0.19	0.20	0.98	0.33
μ_1	7.04	1.09		
μ_2	11.20	1.25		
μ_3	13.23	1.34		
μ_4	17.99	1.55		

All coefficients are positive, as would be expected. *Opportunity*, *Knowledge* and *Concern* do not seem to be related to the measure of effectiveness at conventional levels of significance. The EP students place a higher weight on *Explanation* compared to the EARM students, while the reverse is true for *Enthusiasm.*

There are a number of measures of goodness of fit that can be applied to categorical models. The raw percentage of correct predictions, based on the category with the highest predicted probability, is 0.68, but as is well known, this measure gives a biased impression of the effectiveness of the model, as it does not allow for the underlying distribution of actual responses. The adjusted count R^2 for this model is 0.37, which is relatively high. The McFadden adjusted R^2, (based on the improvement in likelihood value achieved by the model) is 0.34, while the Mckelvey and Zavoina R^2 is 0.69. The

latter is a measure of explained variation, and is an estimate of what the explained sum of squares would be, based on the conditional expectation of the latent variable (Veall and Zimmerman, 1996). Again, this value is relatively high, giving some confidence in the model's ability to explain the factors underlying the measure of *Effectiveness*.

The interpretation of the impact of a change in an explanatory variable within the ordered logit model is not straightforward (Greene, 1997, p. 929). Changes in an exogenous variable will have differential impacts, depending on the category. Table 5 reports the marginal impact of a change in the exogenous variable on the probability of each category (Strongly Disagree, Disagree, etc.) being selected. These changes are evaluated at the mean of the exogenous variables.

Table 5. Marginal Changes in Predicted Probabilities, for Each Category

	Strongly Disagree	Disagree	No Opinion	Agree	Strongly Agree
Explanation (EP)	-0.003	-0.151	-0.309	0.441	0.021
Explanation(EARM)	-0.002	-0.097	-0.200	0.286	0.014
Organzation	-0.001	-0.037	-0.075	0.107	0.005
Opportunity	-0.000	-0.016	-0.034	0.048	0.002
Knowledge	-0.000	-0.025	-0.052	0.074	0.004
Enthusiasm (EP)	-0.001	-0.049	-0.101	0.144	0.007
Enthusiasm (EARM)	-0.002	-0.112	-0.231	0.329	0.016
Concern	-0.000	-0.015	-0.031	0.045	0.002
Predicted prob. using mean values	.0012	0.088	0.339	0.561	0.011

Because of the positive coefficients, increases in the exogenous variables tend to shift the probability distribution to the right, leading to reduced probabilities of the lower categories being selected, and higher probabilities of the upper 2. The differential impact

of *Explanations a*nd *Enthusiasm* between the two courses is now more easily interpreted: a unit increase in *Explanations* raises the probability of selecting Agree for *Effectiveness* by 0.441 for the EP students, but only by 0.286 for the EARM students. Likewise, a unit change in *Enthusiasm* within EARM leads to a 0.329 increase in the probability of selecting Agree for the EARM students, but only 0.144 for the EP students. Although *Organization* is statistically significant, it is noticeable that changes in this variable lead to relatively small changes in the probabilities.

Conclusions

For formative assessment of teaching, an analysis of the determinants of effectiveness may help teachers more efficiently allocate effort to areas that are most important. This may differ for different courses and also between individuals. Knowing what factors have generally been found to be important explanatory variables of effectiveness is of little help to an individual trying to improve their teaching. Most studies of teaching effectiveness examine the determinants of a global effectiveness rating, in effect treating effectiveness as the output in a production function. However, it may be that the relevant question is not how to improve the rating for overall effectiveness but how to make sure that no item rating falls below some trigger point, say 3 on a scale of 1-5.

It is also important to note that we have no information on the relationship between effort expended on a particular component of teaching, and improvement in that item on student evaluations, i.e. the production function for effectiveness. Again, this may vary between individuals in that one person may spend 10 hours to improve the organization of lecture topics with little effect while another may spend the same amount of time and

significantly improve their rating on the *Organization* item, simply due to their inherent abilities. As far as we know, there are no studies on responsiveness of ratings to effort. Professors' time is limited so this information would be a valuable addition to the literature. For example, it may be that certain characteristics are in general more amenable to change than others, although we suspect that this would also vary by individual. While the first author did not record the time spent improving *Explanations,* it was fairly small and resulted in a significant improvement in that rating and thus the overall *Effectiveness* rating.

It is not necessarily true that large coefficients mean that there is more room for improvement as shown by the difference in the mean ratings for *Explanation* and *Enthusiasm*. Given that *Enthusiasm* was already high, little effort was allocated to improving this component of teaching while *Explanations* was quite amenable to change. In addition to ratings, open-ended questions on student evaluations are invaluable as far as providing information as to specific actions that can be taken to improve the course. In addition to student assessments of teaching, assignments and exams can provide valuable input on what concepts need to be explained more clearly in the future. Returning to McKeachie (1975), ratings led to improvement because all three components, new information, motivation, and means, were involved.

The type of reflection and analysis presented can be included in a teaching portfolio. For this purpose, OLS is probably sufficient. Seldin 1999c suggests that adaptations in one's teaching as a result of assessment should be included in self-assessment reports. Seldin (1997 p. 336) points out that "…no matter how effective a particular professor is in the

classroom, he or she can improve. No matter how effective a particular teaching method is, it can be enhanced. These are postulates in higher education."

References

Abrami, P.C. 1989. "How Should we Use Student Ratings to Evaluate Teaching" *Research in Higher Education* Vol. 30(2) pp. 221-227.

Angelo, T.A. and Cross, K.P. 1993. Classroom Assessment Techniques: A Handbook for College Teachers, 2nd Edition. San Francisco: Jossey-Bass, pp. 148-153.

Boex, L.F J. 2000. "Attributes of Effective Economics Instructors: An Analysis of Student Evaluations", *Journal of Economic Education* Vol. 31 (3) pp. 211 – 227.

Cashin, W.E. 1999. "Student Ratings of Teaching: Uses and Misuses" In: P. Seldin and Assoc., (Ed.) Changing Practices in Evaluating Teaching Boston, MA, Anker Publishing Company.

Cohen, P.A. 1981. "Student Ratings of Instruction and Student Achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, Vol. 51, pp. 281-309.

DeCanio, S.J. 1986. "Student Evaluations of Teaching – A Multinomial Logit Approach". *Journal of Economic Education* Vol.17 (3) pp. 165-176.

Eble, K.E. 1988. The Craft of Teaching (2nd Ed.) San Francisco, CA, Jossey-Bass.

Greene, W.H. 1993. Econometric Analysis, 3rd Edition. New Jersey: Prentice-Hall.

Judge, G.C., W.E. Griffiths, R. Carter Hill, H. Lutkepohl, and Tsoung-Chao Lee. 1985. The Theory and Practice of Econometrics. New York, John Wiley and Sons.

McKeachie, W.J. 1975. "Assessing Teaching Effectiveness: Comments and Summary" First International Conference on Improving University Teaching, Heidelberg, Germany. (as cited in Seldin 1997).

McKeachie, W.J. 1986. Teaching Tips: A Guidebook for the Beginning College Teacher (8th Ed.) Lexington, MA, D.C. Heath and Co.

Seldin, P. 1997. "Using Student Feedback to Improve Teaching", In: D. DeZure (Ed.) *To Improve the Academy*, Vol. 16, pp. 335-346. Stillwater, OK, New Forums Press.

Seldin, P. 1999a. "Chapter 1. Current Practices – Good and Bad – Nationally" In: P. Seldin and Assoc., (Ed.) Changing Practices in Evaluating Teaching Boston, MA, Anker Publishing Company.

Seldin, P. 1999b. "Preface" In: P. Seldin and Assoc., (Ed.) Changing Practices in Evaluating Teaching Boston, MA, Anker Publishing Company.

Seldin, P. 1999c. “Chapter 5. Self-Evaluation: What Works? What Doesn’t?” In: P. Seldin and Assoc., (Ed.) Changing Practices in Evaluating Teaching Boston, MA, Anker Publishing Company.

StataCorp. 1999. Stata Statistical Software: Release 6.0. College Station, TX: Stata Corporation.

Ory, J.C. 2000. “Teaching Evaluation: Past, Present and Future” *New Directions in Teaching and Learning*, No. 83, Fall 2000.

Veall, M.R. and Zimmermann,K.F. 1996 “Psuedo-R^2 measures for some common limited dependent variable models” *Journal of Economic Surveys*, Vol.10 (3) pp. 241-259.

Zubizarreta, J. 1999. “Chapter 9. Evaluating Teaching Through Portfolios” In: P. Seldin and Assoc., (Ed.) Changing Practices in Evaluating Teaching Boston, MA, Anker Publishing Company.