FEA Working Paper No. 2002-3

BIASES IN STUDENT EVALUATION OF TEACHING: THE CASE OF FACULTY OF ECONOMICS & ADMINISTRATION, UNIVERSITY OF MALAYA

Liaw Shu Hui*
Goh Kim Leng**

Department of Applied Statistics
Faculty of Economics & Administration
University of Malaya
50603 Kuala Lumpur
MALAYSIA

email: *liawsh@um.edu.my
**klgoh@um.edu.my

May 2002

Biases in Student Evaluation of Teaching: The Case of Faculty of Economics & Administration, University of Malaya

Abstract

This paper shows that class size has inappropriately influenced students' judgements on evaluation of lecturers in the Faculty of Economics and Administration, University of Malaya. A bias exists whereby courses with small enrolment receive good overall teaching ratings, whereas larger classes have produced poor evaluations. On the other hand, teaching ratings are not affected by instructor characteristics (instructional experience, rank and gender) or course characteristics (type and level of subject, and time and day course is taught). To improve the construct validity of student ratings as a measure of teaching performance, this paper suggests using results from regression analysis to estimate the bias factor associated with class size, and adjusting the overall teaching ratings based on this estimate in order to remove the class size bias.

Introduction

Student evaluation of teaching (SET) is commonly used in tertiary institutions to measure teaching performance, with the prime purpose of improving course and teaching quality. This tool is also often used as part of the evaluation process for staff appraisal. The reliability and validity of SET as a measure of teaching effectiveness has been the subject of much research. The review by Greenwald (1997) shows that there are more researchers who advocated the validity of SET than those who viewed this otherwise.

However, the acceptance of its validity does not mean that all SET instruments are reliable. The reliability and usefulness of the instruments depend on the content and items being measured. A lack of understanding of what is actually being measured by the SET instrument often leads to misuse and criticism. Bedggood and Pollard (1999), for example, argued that SET can be seriously flawed when used as the sole measure of teaching effectiveness based on the outcome of their examination of the survey design of SET in eight Australian universities. It is generally recognized that SET should be considered concurrently with other indicators of effective teaching such as retrospective evaluations by alumni, teachers' self-evaluations, observations by trained experts or the quality of students' learning, and this is particularly true if the purpose is for personnel evaluation decisions (Peterson et al. 1998; Marsh and Roche 1997).

Teaching is a complex activity with multiple dimensions. SET instruments should reflect this multidimensionality. However, there is no general consensus as to the use of global overall ratings for summative evaluation, multiple ratings or weighted average of specific factors of SET measurements (Marsh 1987, 1994; Cashin and Downey 1992; Cashin, Downey and Sixbury 1994; Marsh and Roche 1993, 1997; Apollonia and Abrami 1997). SET is useful as a teaching evaluation method, but only to the extent that ratings objectively reflect the qualities they are designed to measure. Unfortunately, many empirical studies indicated that SET measures not only aspects of instructional effectiveness, but captures also factors that are normatively irrelevant to teaching quality. The latter introduces biases in the results of SET and hence reduces its creditability as a measure of teaching quality. Among others, evidence of such biases are found in the studies by Marsh and Overall (1981), Basow and Silberg (1987), Goldberg and Callahan (1991), and Langbein (1994) for the case of U.S.; Marsh and Roche (1993) for Australia; Husbands (1997) for U.K.; and Koh and Tan (1997) for Singapore. A section of Marsh and Roche's (1997) work details further evidence. Although many of these studies included samples covering different schools/faculties, and some even universities, a significant number focussed on the SET of one school or faculty (for example, Cranton and Smith 1986; Langbein 1994; Koh and Tan 1997).

This study examines the results of the first SET conducted by the Faculty of Economics and Administration (FEA), University of Malaya in 1998 after more than 30 years since its inception, in order to determine if student ratings are influenced by factors which are not directly relevant to the quality of instruction. This is motivated by the objective of the implementation of SET at FEA. It was maintained that the SET was conducted to obtain feedback from students for course and teaching improvement. Although not in the spirit in which the SET was initiated, the ratings were inevitably utilized as a measure of teaching performance. A good understanding of the factors apart from teaching performance that affect SET is thus important to avoid misuse of this instrument. Further, this paper provides an example as to how such bias factors can be quantified and subsequently removed from the SET ratings.

The 1998 SET is of special interest as for this particular exercise, the faculty undertook full responsibility in designing the questionnaires, implementing the survey, processing the data and computing the survey results. A year after, the SET was centralized and the Quality Assurance Unit of University of Malaya spearheaded a university wide exercise. The involvement of the faculty was mainly restricted to distributing and collecting the survey forms, while the other tasks went under the jurisdiction of the Quality Assurance Unit. At least for the year in which the FEA had full autonomy in conducting the SET, a consensus is reached among the faculty members on questions to be included in the student opinion survey. This stands in contrast to a university wide survey that is so plagued by the need to adjust questionnaires to the different teaching and learning objectives of vastly different disciplines.

The organization of this paper is as follows. Following the introduction, the next section reviews works that investigated factors affecting class evaluation ratings in order to establish the important determinants to be used for the analysis in this study. The third section describes the data and methodology employed. The findings are reported in the section after. This section also provides an example as to how biases can be removed from the ratings of teaching evaluation. The final section summarizes the findings and discusses the implications.

Literature Review

Some variables that are perceived to have an influence on outcome of SET include class size; the level and type of subject; day and time of the course; gender, rank and instructional experience of the instructor; and the response rate of SET. The results of research on these potential bias factors are mixed. The findings seem to be dependent on context, content as well as methodology of teaching evaluation. Despite this, however, the point that stands out strongly is the presence of biases. The only question that remains is the type and extent of biases.

A comprehensive survey of earlier works on the relationship between class size and student ratings can be found in Feldman (1984). Generally, there is a consistent trend for ratings to increase for smaller classes. A secondary pattern is the U-shaped relationship between class size and teaching evaluation ratings, where instructors of relatively small and relatively large classes received better ratings. There might also be a complex relationship between class size and teaching evaluation ratings interacting with other factors such as the course level and major field of study, as reported by Cranton and Smith (1986). On the other hand, some studies found only a weak or no significant relationship between class size and overall ratings of course and instructor (Marsh 1987, Langbein 1994, Koh and Tan 1997).

Subjects taught at higher levels often received better SET ratings. Aleamoni and Hexner (1980) cited 18 articles that supported this finding. More recent studies reporting a positive association between global evaluation ratings and course level include those of Goldberg and Callahan (1991) and Langbein (1994). There were still others showing that instructors of middle-level subjects received relatively poorer ratings or the level of subjects was confounded with other factors (Cranton and Smith 1986; Koh and Tan 1997). On the contrary, Aleamoni and Hexner (1980) cited eight studies that found no relationship between course level and ratings.

The type of courses taken by students can be broadly divided into two categories, namely, quantitative and non-quantitative. The method of teaching varies grossly across these categories. There is a tendency to perceive that quantitative subjects usually receive lower ratings than non-quantitative subjects (Langbein 1994). Marsh and Overall (1981) indicated that the effect of the type and level of course were relatively less important than the effect of instructor who teaches it. In addition, Langbein (1994) found that there was no significant relationship between the type of course and overall instruction ratings.

Some researchers suggested that the time and day a course is taught can affect SET results (DeBerg and Wilson 1990; Husbands and Fosh 1993). For courses scheduled to be held late in the day, students' perception of course effectiveness could be discounted by their tiredness after a long day. However, there are studies that also indicated that the time schedule of a course does not relate significantly to overall teaching ratings (Aleamoni 1981; Koh and Tan 1997).

Besides course characteristics, teacher characteristics are also believed to have an influence on teaching evaluation ratings. A great deal of attention has been devoted to study the effect of gender on student ratings of instructor. Basow and Silberg (1987), for example, examined specifically if male and female instructors are rated differently and reviewed many related works. As college teaching has been traditionally viewed as a male-dominated profession, it is generally found that female teachers are being rated more negatively than their male counterparts. The same finding is reported by Langbein (1994). Freeman (1994) suggested that women's roles in academia have changed and his findings indicate that gender-typed characteristics are more important in affecting student evaluations than is instructor gender itself. Characteristics including those typified as masculine (e.g., assertive and forceful) and as feminine (e.g., affectionate and sensitive), are both deemed important by students.

Findings from previous studies are rather mixed on the relationship between the academic rank, instructional experience or age of an instructor and the overall evaluation of his/her teaching. Close to sixty studies were reviewed by Feldman (1983). About half found no significant association between the measure of seniority and SET. Half of the remaining studies reported a significant positive relationship, while the other half indicated a significant negative association. Langbein (1994), on the other hand, found that instructional experience has a significant nonlinear relationship with overall instruction ratings. Evaluations become more positive as years of teaching experience increase, but after the mid-teen years of experience, more experience turns student evaluations in a negative direction.

Studies on effect of response rate of SET are relatively scarce in the literature. Koh and Tan (1997) found that a large number of evaluation responses resulted in better overall ratings. To this, they conjectured that the likelihood is high that students who are motivated to respond to a teaching evaluation are those that show interest in the subject.

Data and Methodology

The data used in this study are information collected from the teaching evaluation exercise for the first and second semester of the 1998/99 academic session. In the last two weeks of these semesters, responses were elicited from the students using a questionnaire. This study focusses on the responses for the bachelor degree programmes, and the instructors were evaluated separately on lectures conducted throughout the semester for each of the courses they taught. The questionnaire for the first semester had seven statements, and eight were included in the questionnaire for the second semester. The responses to five statements that are common for both semesters and relate directly to teaching are used for analysis. These statements are:

1. Organization: The lecturer plans each lecture in detail and systematically.
2. Knowledge: The lecturer is knowledgeable in the subject taught.
3. Presentation: The lecturer is able to present the lecture in an interesting way.
4. Clarity: The lecturer is able to deliver the lecture clearly.
5. Appropriateness: The lecturer gives appropriate examples to illustrate the subject matter.

Each statement was rated on a 5-point scale ranging from disagree strongly (1) to agree strongly (5). Based on the scores for these five statements, an overall teaching rating is obtained for every lecturer evaluated. The mean for each statement is firstly computed by weighting the score with the corresponding proportion of responses to this score. The percentage of this weighted mean of the highest possible attainable score is then taken. The mean of the percentages for the five statements are used to represent the overall teaching rating.

As a preliminary analysis, a univariate approach is used to investigate if potential biases associated with course and instructor characteristics exist. The F-test and Kruskal-Wallis test for equality of group means are adopted to examine if the overall teaching ratings are different within these characteristics. The course characteristics are the level of subject (first, second or third year), type of subject (qualitative or quantitative), time of the lecture (morning or afternoon) and day of the lecture (Monday to Friday). The instructor characteristics are gender and rank (lecturer, associate professor or professor).

In order to analyze the controlled effects of the course and instructor characteristics on the overall teaching ratings, the multiple regression is used. In addition to the variables mentioned, the size of enrolment for the course, instructional experience measured as the number of years of service at the FEA, and the response rate of SET for each course are also included. The model is as follows:

$$\begin{aligned} Y_i = {} & \alpha_0 + \alpha_1(\text{male}_i) + \alpha_2(\text{prof}_i) + \alpha_3(\text{assoc}_i) + \alpha_4(\text{year1}_i) + \alpha_5(\text{year2}_i) + \alpha_6(\text{quan}_i) \\ & + \alpha_7(\text{morn}_i) + \alpha_8(\text{mon}_i) + \alpha_9(\text{tues}_i) + \alpha_{10}(\text{wed}_i) + \alpha_{11}(\text{thur}_i) \\ & + \alpha_{12}(\text{exp}_i) + \alpha_{13}(\text{size}_i) + \alpha_{14}(\text{resp}_i) + \varepsilon_i \qquad \text{..................} \ (1) \end{aligned}$$

where

Y = overall teaching rating
male = 1 if instructor is male, 0 otherwise.
prof = 1 if instructor is a professor, 0 otherwise.
assoc = 1 if instructor is an associate professor, 0 otherwise.
year1 = 1 if course is first year level, 0 otherwise.
year2 = 1 if course is second year level, 0 otherwise.
quan = 1 if course is quantitative, 0 otherwise.
morn = 1 if lecture is held in the morning, 0 otherwise.
mon = 1 if lecture is held on Monday, 0 otherwise
tues = 1 if lecture is held on Tuesday, 0 otherwise
wed = 1 if lecture is held on Wednesday, 0 otherwise
thur = 1 if lecture is held on Thursday, 0 otherwise
exp = instructional experience (years)
size = class size (number of students)
resp = response rate (percent)
ε = error term

In classical regression models, the error term should be identically and independently distributed. However, out of a total of 48 lecturers evaluated, 19 taught only one course, 20 taught two courses, 8 taught 3 courses and 1 taught 4 courses. An examination of the overall teaching ratings for the lecturers who taught at least two courses reveals that the same instructor receives very similar scores.[1] This violates the assumption of independence, and thus, only one rating is selected randomly to be included in the analysis for the lecturers teaching two or more courses. To deal with estimation problems associated with changing variance in the error term, the White heteroscedasticity-consistent standard errors (White 1980) are used in the computation of the t-statistics for tests of significance.

The presence of outliers is checked. The overall rating of all the lecturers is at least 65 percent and above, except one, who received a rating below 50 percent. This is the only case outside the 3 standard error bound and is excluded from the analysis.

Analysis and Results

The instructional experience of lecturers ranges from 2 to 29 years with a mean of 12.67 years and a standard deviation of 7.97 years. The class size ranges from 10 to 301, with a mean of 85.8 and standard deviation of 62.7. The distribution of class size is heavily skewed to the right, and 15 classes have enrolment of more than 100 students. The response rate on teaching evaluation has a mean of 64.88 percent and varies widely with a standard deviation of 21.58 percent. The statistics related to the course and lecturer characteristics are reported in Table 1. At the first year level, only compulsory courses are offered and therefore the number is smaller compared to the other two levels. Most of the lecturers have a preference to schedule their teaching in the morning. The Friday afternoons were kept free from teaching for the faculty meetings and seminars, and this resulted in a smaller number of evaluations for Friday.

[1] The range of the overall ratings for two lecturers (14.6 percent and 10.2 percent) is much higher than the others (well below 10.0 percent). These lecturers were also part-time PhD candidates. The double commitment of teaching and studying might have contributed to the inconsistency in their teaching. These two cases are excluded from the study.

Table 1
Number of Cases and Summary Statistics of Overall Teaching Rating
by Course and Instructor Characteristics

Variable	Number of cases	Percent	Overall Teaching Rating		
			Mean	Median	Std. deviation
Gender of Instructor					
Male	23	51.1	80.36	80.80	8.14
Female	22	48.9	77.90	77.95	9.22
Rank of Instructor					
Lecturer	30	66.7	79.20	79.40	8.59
Associate Professor	8	17.7	81.49	82.35	7.11
Professor	7	15.6	76.30	79.00	10.96
Level of Course					
First year	9	20.0	77.82	76.80	5.80
Second year	13	28.9	79.28	84.20	10.44
Third year	23	51.1	79.61	79.50	8.83
Type of Course					
Qualitative	26	57.8	78.33	79.25	9.76
Quantitative	19	42.2	80.29	79.60	7.03
Time of Lecture					
Morning	31	68.9	78.97	79.50	9.13
Afternoon	14	31.1	79.58	78.90	7.89
Day of Lecture					
Monday	11	24.4	79.23	83.40	10.61
Tuesday	8	17.8	77.63	76.20	9.83
Wednesday	12	26.7	79.93	79.25	6.28
Thursday	8	17.8	77.55	76.85	7.63
Friday	6	13.3	81.67	85.30	10.81
Total	45	100.0	79.16	79.50	8.68

The mean of the overall teaching rating is 79.2 percent, close to the median of 79.5 percent. Table 1 also shows the statistics for different groups of gender and rank of lecturers, level and type of courses, and the time and day a course is scheduled. It can be seen that the group means for all the variables are rather close. The F-test and Kruskal-Wallis test were performed to examine if there are group differences within a variable and the results are given in Table 2. In all the cases, no significant differences are observed. This univariate analysis suggests that the course and lecturer characteristics considered are not potential sources of bias to the results of SET.

Table 2
The results of F-test and Kruskal-Wallis Test for Differences in Group Means
for Course and Instructor Characteristics

Variable	F-test	Kruskal-Wallis Test
	(p-value)	(p-value)
Gender of Instructor	.348	.340
Rank of Instructor	.523	.739
Level of Course	.875	.625
Type of Course	.460	.696
Time of Lecture	.829	.825
Day of Lecture	.901	.472

Note: See Table 1 for grouping within each variable.

The full model specified in equation (1) is estimated and the results are shown in Table 3. Only the response rate is significant at the 5 percent level. Year of experience and class size does not appear to be significant. The other variables that are not significant include the instructor characteristics (gender and rank) and the course characteristics (level and type of subject, and time and day of lecture). This confirms the earlier findings that the course and instructor characteristics are not sources of bias to the SET results. A low R^2 is expected, as the overall teaching index should best not be explained by the variation in the potential bias factors (see Koh and Tan 1997).

The model is passed through a battery of diagnostic tests – the Jarque-Bera test of normality (Gujarati 1995, p. 143), White test of heteroscedasticity (Gujarati 1995, p. 379) and Ramsey test of regression specification (Gujarati 1995, p. 464). The model does not suffer from problems of departure from normality, heteroscedasticity and omission of important variables. The problem of multicollinearity, however, is serious. A series of auxiliary regressions are estimated to examine the relationships between the independent variables and Table 4 reports the F-test for the significance of these relationships. As expected, the number of years of experience is significantly related to the rank of the lecturer, as those who have been promoted are more likely to be the more senior staff. The level of courses is related to the class size because the first year courses are compulsory and hence have large enrolments. Since no teaching is scheduled on Friday afternoons, the time of lecture is systematically related to the day of lecture. Also, a significant relationship is found between the response rate and class size. Interestingly, these two variables are negatively related (Pearson coefficient of correlation = –0.375, significant at 5 per cent), indicating that attendance when the teaching evaluation was conducted is poorer for larger class size.

Due to multicollinearity, a smaller set of explanatory variables is used in the regression model, with the following omitted from the regression – prof, assoc, year1, year2 and morn. The marginal contribution of these variables is also not significant (F-statistic = 2.92, p-value = 0.71). The response rate, which is significant, is retained in the new model and the results are given in Table 5. At the 5 percent level, response rate remains to be the only significant explanatory variable. The higher the response rate, the better is the overall teaching rating. The results can be viewed in two ways. If the class attendance on the day when the evaluation was conducted is representative of the usual attendance, an ineffective lecturer may experience typically poor turn out for his/her course. Thus, student learning may exert its influence on class attendance and subsequently the rating of the lecturer.

Table 3
The Full Regression Equation for Explaining the Overall Teaching Rating

Variable	Coefficient	Std. Error	t-statistic	p-value
constant	78.48**	6.90	11.38	0.00
male	3.04	2.64	1.15	0.26
prof	-0.98	6.99	-0.14	0.89
assoc	2.92	4.52	0.65	0.52
year1	0.72	3.95	0.18	0.86
year2	-3.37	3.81	-0.88	0.38
quan	4.22	3.07	1.38	0.18
morn	-1.21	3.10	-0.39	0.70
mon	-5.34	4.91	-1.09	0.29
tues	-5.50	5.09	-1.08	0.29
wed	-3.57	4.76	-0.75	0.46
thur	-9.07	5.28	-1.72	0.10
exp	-0.18	0.26	-0.71	0.48
size	-0.03	0.02	-1.32	0.20
resp	0.13*	0.06	2.19	0.04

$R^2 = 0.32$
Adjusted $R^2 = 0.01$

White test for heteroscedasticity:	Test statistic = 23.00	p-value = 0.15
Ramsey RESET test:	Test statistic = 0.62	p-value = 0.43
Jarque-Bera test of normality:	Test statistic = 2.02	p-value = 0.36

Notes: The dependent variable is overall teaching rating.
See text for definition of the variables.
The White test includes squared but not cross terms.
The Ramsey test includes only one fitted term in the test regression.
**Significant at 1 percent.
*Significant at 5 percent.

Table 4
The Significance of Auxiliary Regressions for Testing Presence of Multicollinearity

Dependent Variable	Independent Variables	F-statistic (p-value)
exp	prof, assoc	26.53 (0.00)
size	year1, year2	7.60 (0.00)
morn	mon, tues, wed, thur	3.92 (0.01)
resp	size	7.04 (0.01)

Notes: The F-statistic is for testing the overall significance of the regression. See text for definition of the variables.

Table 5
Regression Equation for Explaining the Overall Teaching Rating
with Response Rate as an Explanatory Variable

Variable	Coefficient	Std. Error	t-statistic	p-value
constant	72.79**	5.33	13.65	0.00
male	3.29	2.38	1.38	0.18
quan	3.50	2.19	1.60	0.12
mon	-4.50	4.84	-0.93	0.36
tues	-5.67	4.75	-1.19	0.24
wed	-3.00	4.05	-0.74	0.46
thur	-7.72	4.38	-1.76	0.09
exp	-0.15	0.17	-0.88	0.39
resp	0.15**	0.05	2.73	0.01

$R^2 = 0.24$
Adjusted $R^2 = 0.07$

White test for heteroscedasticity:	Test statistic = 7.51	p-value = 0.68
Ramsey specification test:	Test statistic = 0.54	p-value = 0.47
Jarque-Bera test of normality:	Test statistic = 2.46	p-value = 0.29

Notes: See notes to Table 3.

Given that response rate is significantly related to class size, the other view is that class size is a bias factor. This is confirmed by estimating the regression with class size replacing response rate and the results are reported in Table 6. The problem of multicollinearity renders this variable insignificant in the original model, but class size is significant at the 5 percent level in the new model. The negative coefficient shows that the overall teaching rating drops as class size increases. This is expected as a larger class size reduces the lecturer's interaction and interrelationships with students, and hence could bias the student in their evaluation.

The regression results show that the bias factor associated to class size is –0.05 for every additional student enrolled for the course. This factor is used for making adjustment to the overall teaching ratings. The enrolment for all the courses is assumed to be 85.8 students, which is the mean size of all the classes. The use of the mean class size as control ensures that the mean of the overall teaching ratings is maintained after the adjustment is made. The new ratings are obtained by taking away the total bias computed for the difference between the actual enrolment and the mean class size from the old ratings. This means that the overall teaching rating is adjusted upwards for courses with class size above the average, and downwards for those with class size below the average.

To evaluate how effective this method is, the old and new ratings are ranked, where the highest overall teaching rating is given a rank of 1 in both cases. Table 7 shows the correlation coefficients between the ratings, ranks and class size. Consistent with the implications of the regression results, the ratings before adjustment and the corresponding ranks are both significantly correlated with class size. However, this significance in correlation disappears after the adjustment to the overall teaching ratings is made. This indicates that the class size bias factor is successfully removed.

Table 6
Regression Equation for Explaining the Overall Teaching Rating
with Class Size as an Explanatory Variable

Variable	Coefficient	Std. Error	t-statistic	p-value
Constant	85.57**	5.54	15.45	0.00
Male	2.72	2.42	1.13	0.27
Quan	3.77	2.08	1.81	0.08
Mon	-3.34	5.36	-0.62	0.54
Tues	-3.40	5.12	-0.66	0.51
Wed	-1.25	4.52	-0.28	0.78
thur	-6.41	4.64	-1.38	0.18
exp	-0.21	0.18	-1.17	0.25
size	-0.05*	0.02	-2.28	0.03

R^2 = 0.21
Adjusted R^2 = 0.03

White test for heteroscedasticity:	Test statistic = 10.33	p-value = 0.41
Ramsey specification test:	Test statistic = 0.40	p-value = 0.53
Jarque-Bera test of normality:	Test statistic = 3.06	p-value = 0.22

Notes: See notes to Table 3.

For the adjustment to be meaningful, it should not totally distort the pattern of the original ranking of the overall teaching ratings. This is important as although class size introduces bias, the adjustment should not penalize the effective lecturers who obtained high ratings on their own merit, nor should it benefit those with low ratings due to poor teaching skills. The correlation coefficients between the old and new ratings, and the old and new rankings, are still about 0.90 and highly significant in the right direction. This indicates that some form of stability is maintained while the class size bias is removed. Figure 1 plots the old and new rankings. The ranking of the top 5 highest ratings, and the bottom 5 lowest ratings, does not change much. Overall, there is no change in the ranking of 20% of the lecturers, while 20% of them experienced a change of only one rank. The runs test is performed to check if a drop or increase in rank after the adjustment for class size bias is dependent on the original ranking. The test does not reject the randomness in the sign of the differences between the original and new ranks arranged in ascending order of the original ranking (Z = 1.001, p-value = 0.317). This suggests that the adjustment process is in no way systematically linked to the original ranking, and therefore, does not necessarily penalize those with high ratings or benefit those with low ratings.

Table 7
Correlation Matrix for Ranking, Overall Teaching Ratings and Class Size

Pearson Correlation Coefficient

	New ranking	New ratings	Original ranking	Original ratings	Class size
New ranking	1.000	-0.936**	0.898**	-0.886**	-0.015
New ratings		1.000	-0.874**	0.945**	0.024
Original ranking			1.000	-0.955	0.374*
Original ratings				1.000	-0.305*
Class size					1.000

Spearman’s Rank Correlation Coefficient

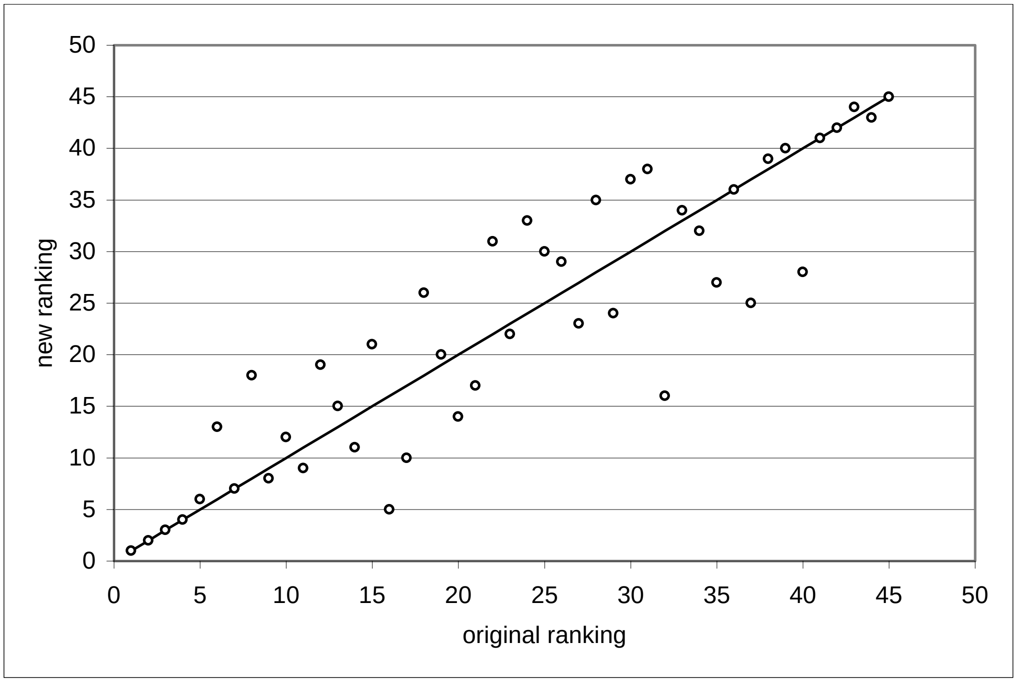
	New ranking	New ratings	Original ranking	Original Ratings	Class size
New ranking	1.000	-1.000**	0.898**	-0.900**	0.032
New ratings		1.000	-0.898**	0.900**	-0.032
Original ranking			1.000	-1.000**	0.391**
Original ratings				1.000	-0.389**
Class size					1.000

Notes: **Significant at 1 percent.
* Significant at 5 percent.

Conclusion

The main finding of this study is that class size has a statistically significant direct effect on the evaluation of teaching performance of the lecturers in FEA. Courses with small enrolment receive good overall teaching ratings, whereas larger classes have produced poor evaluations. This would necessarily be an indicator that student ratings are biased by class size. Recognizing the fact, this study has used the results from a multiple regression to estimate the bias factor associated with class size and demonstrated how this estimate can be used to remove the bias.

Figure 1
Ranking of the Overall Teaching Ratings Before and After Adjusting
for the Class Size Bias Factor



The ability to remove this bias will improve the construct validity of student surveys, but does not immediately suggest that student ratings of teaching can be relied upon exclusively as a measure of teaching effectiveness. It must be clear that the overall teaching rating obtained from the FEA student survey is only a measure of the instructor's organization and presentation skills and therefore does not reflect entirely the quality of teaching or how much the student is learning, and hence, must be used cautiously.

Undoubtedly, lecturers who can motivate students to work hard deserve a favourable evaluation, and to this extent, the student ratings reflect at least some of what quality teaching is supposed to do. Teaching effectiveness could be enhanced if the learning environment is made more conducive by keeping the class size small, as is consistent with the finding that smaller classes get better evaluations. This highlights the importance of low student-staff ratio, particularly for promoting group interaction in classes and to enable the instructor to play the facilitative role effectively. Further, a smaller class size helps to develop better interrelationship between the instructor and students, and the instructor also commands better attention from the students. These factors would necessary have an indirect positive effect on the rating of the instructor. To this extent, student evaluation of teaching can only be a more effective tool for measuring teaching performance if variability in class size within the faculty is reduced.

References

Aleamoni, L.M., 1981. *Student ratings of Instruction* in Millman, J. (Ed.), *Handbook of Teacher Evaluation*, Sage Publications, CA, 110-145.

Aleamoni, L.M. and Hexner, P.Z., 1980. A Review of the Research on Student Evaluation and a Report on the Effect of Different Sets of Instructions on Student Course and Instructor Evaluation, *Instructional Science*, 9: 67-84.

Basow, S.A. and Silberg, N.T., 1987. Student Evaluations of College Professors: Are Female and Male Professors Rated Differently? *Journal of Educational Psychology*, 79(3): 308-314.

Bedggood, R.E. and Pollard, R.J., 1999. Uses and Misuses of Student Opinion Surveys in Eight Australian Universities. *Australian Journal of Education*, 43(2): 129-141.

Cashin, W.E. and Downey R.G., 1992. Using Global Student Rating Items for Summative Evaluation. *Journal of Educational Psychology*, 84(4): 563-572.

Cashin, W.E., Downey R.G. and Sixbury, G.R., 1994. Global and Specific Ratings of Teaching Effectiveness and Their Relation to Course Objectives: Reply to Marsh (1994). *Journal of Educational Psychology*, 86(4): 649-657.

Cranton, P.A. and Smith, R.A., 1986. A New Look at the Effect of Course Characteristics on Student Ratings of Instruction. *American Educational Research Journal*, 23(1): 117-128.

D'Apollonia, S. and Abrami, P.C., Navigating Student Ratings of Instruction. *American Psychologist*, 52(11), 1198-1208.

DeBerg, C.L. and Wilson, J.R., 1990. An Empirical Investigation of the Potential Confounding Variables in Student Evaluation of Teaching. *Journal of Accounting Education*, 8(1), 37-62.

Feldman, K.A., 1983. Seniority and Experience of College Teachers as Related to Evaluations They Receive from Students. *Research in Higher Education*, 18(1): 3-124.

Feldman, K.A., 1984. Class Size and College Students' Evaluations of Teachers and Courses: A Closer Look. *Research in Higher Education*, 21(1): 45-116.

Freeman, H.R., 1994. Student Evaluations of College Instructors: Effects of Type of Course Taught, Instructor Gender and Gender Role, and Student Gender. *Journal of Educational Psychology*, 86(4): 627-630.

Goldberg, G. and Callahan, J., 1991. Objectivity of Student Evaluations of Instructors. *Journal of Education for Business*, 66(6): 377-379.

Greenwald, A.G., 1997. Validity Concerns and Usefulness of Student Ratings of Instruction. *American Psychologist*, 52(11): 1182-1186.

Gujarati, D.N., 1995. *Basic Econometrics*, 3rd edition, McGraw-Hill, New York.

Husbands, C.T. and Fosh, P., 1993. Students' Evaluation of Teaching in Higher Education: Experiences from Four European Countries and Some Implications of the Practice. *Assessment and Evaluation in Higher Education*, 18(2), 95-114.

Koh, H.C. and Tan, T.M., 1997. Empirical Investigation of the Factors Affecting SET results. *International Journal of Educational Management*, 11(4): 170-178.

Langbein, L.I., 1994. The Validity of Student Evaluations of Teaching. *Political Science and Politics*, 27(3): 545-553

Marsh, H.W., 1987. Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research. *International Journal of Educational Research*, 11, 253-388.

Marsh, H.W., 1994. Comments on Weighting for the Right Criteria in the Instructional Development and Effectiveness Assessment (IDEA) System: Global and Specific

Ratings of Teaching Effectiveness and Their Relation to Course Objectives. *Journal of Educational Psychology*, 86(4): 631-648.

Marsh, H.W. and Roche, L.A., 1993. The Use of Students' Evaluations and an Individually Structured Intervention to Enhance University Teaching Effectiveness. *American Educational Research Journal*, 30(1): 217-251.

Marsh, H.W. and Roche, L.A., 1997. Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias and Utility. *American Psychologist*, 52 (11): 1187-1197.

Marsh, H.W. and Overall J.U., 1981. The Relative Influence of Course Level, Course Type, and Instructor on Students' Evaluations of College Teaching. *American Educational Research Journal*, 18(1): 103-112.

Peterson, K.D., Stevens D. and Ponzio R.C., 1998. Variable Data Sources in Teacher Evaluation. *Journal of Research and Development in Education*, 31(3):123-132.

White, H., 1980. A heterocedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity, *Econometrica*, 48: 817-838.