The Canadian Journal of Higher Education, Vol. XI-2, 1981
La revue canadienne d'enseignement supérieur, Vol. XI-2, 1981

Student Evaluation of Teaching[1]

DOUGLAS J. McCREADY*

ABSTRACT

The analysis of factors which influence student valuations of teaching is the subject of this paper. An empirical test, using the evaluations carried out in the School of Business and Economics at Wilfrid Laurier University shows that the grades granted by instructors do not relate significantly to the evaluations of that instructor by students. Factors which do relate to higher evaluations include: early morning classes, small classes, optional subjects, and senior classes. From a survey of how faculty react to the evaluations, it appears that most faculty do not find the evaluations useful in making improvements in their own teaching.

RESUME

La présente communication a pour but d'analyser les facteurs qui influencent les étudiants dans leur évaluation des professeurs. Une étude, basée sur des sondages effectués au School of Business and Economics de Wilfrid Laurier University, montre que l'opinion des étudiants ne semble pas proportionnelle aux notes attribuées par les professeurs. Les évaluations positives sont reliées aux facteurs suivants: heure matinale du cours, petites classes, matières facultatives et cours avancés. Un autre sondage semble montrer que la plupart des professeurs ne trouvent pas les évaluations très utiles pour perfectionner leur enseignement.

INTRODUCTION

Teaching evaluation can be a highly contentious issue amongst the professoriat. This is particularly so when the main purpose of teacher and/or course evaluation by students is to provide a basis for administrative decisions relating to tenure, promotion, and salary. Those who dispute the value of students' evaluation of faculty contend that there are a number of potential sources of error in the instruments used to evaluate teaching in most universities, including:

1. the possibility of an implicit or explicit contract between the professor and student to give high evaluations to each other;
2. the possibility that tests, assignments and return dates thereof can influence the evaluation;

* School of Business and Economics Wilfrid Laurier University.

3. the likelihood that attendance in the weakest classes will be low on evaluation day thus allowing weak professors to do better on the evaluation than mediocre professors where attendance at class is large;
4. the probability that professors who teach compulsory courses have to overcome a natural negative bias while faculty teaching option courses have students who are positively predisposed to the professor;
5. the fact that students react to courses taught in time slots that interfere with "more enjoyable activities".

The purpose of this note is to review some of these potential biases using the teaching evaluation form used by the School of Business and Economics at Wilfrid Laurier University for empirical data.

It is obvious that a survey of the students on the value of a given instructor are potentially subject to a number of potential influences. It is these which are to be evaluated here.

Whether students tend to rate courses more highly when they receive or expect good grades in them is a much debated topic in the literature. Does the evidence support an assertion that a teacher can get "good" ratings simply by assigning "good" grades? Some studies appear to demonstrate just that. In fact, Anikeef (1953), Brown (1976), Caffrey (1969), Jiobu and Pollis (1971), Perry and Baumann (1973), Pohlmann (1975), and Rayder (1968) amongst others report significant positive relationships between students' grades and their ratings of instructors. However, the correlation coefficient in these studies, as pointed out by Costin, Greenough, and Mengers (1971), is typically, although not always, smaller than .30. In contrast, no statistically significant relationship between average class grades and student evaluations of the instructor were found in studies conducted by Sherman and Blackburn (1975) and Voecks and French (1960).

Positive correlations between grades and student course evaluations are sometimes decreased in value by other factors. For example, Anikeef (1953) finds the relationship to be less strong for senior students than for freshmen. Feldman (1976) correctly points out that if course (structural) characteristics relate to both average grades and average evaluation, these characteristics need to be incorporated in the analyses and interpretations, if the meaning of the relationship between grades and evaluation is not to be ambiguous.

In a study of 2,750 student appraisal forms at Baldwin-Wallace College, S. Lee Whiteman found:[2]

> No significant relationship between student's reported grade point average and his ratings on the form
>
> ($r = +.01$, not significant at .01 level)

Whiteman has data on size of class, year in which the course is normally taken, and how the course is timetabled.

Murray (1980) reports on a significant number of studies in which student evaluations of instructors are correlated with grades, and with objective measures of student learning. He also reports on the "Dr. Fox effect" in which an actor who says next to nothing in an enthusiastic way is accorded a high rating. Murray concludes that since most studies find a moderate positive relationship between objective measures of learning and student ratings of global performance (but not as good a relationship with rapport or feedback), there is some validity in using student ratings of overall effectiveness of the instructor in administrative decisions, regardless of the positive correlation between grades and student evaluations.

TABLE 1

Summary of Data Used

Term	Number of Course Registrants (1)	Number of Offerings (2)	Number of Valid Observations (3)
		BUSINESS[a]	
Fall and Winter, 1975-76	8,951	267	104
Intersession & Summer, 1976	1,352	29	0
Fall and Winter, 1976-77	9,256	278	140
		ECONOMICS[b]	
Fall and Winter, 1975-76	3,031	48	38
Intersession & Summer, 1976	451	10	0
Fall and Winter, 1976-77	3,349	40	36
TOTAL	26,390	672	318

[a]All business courses are one-term courses.

[b]In 1975-76, 28 two-term offerings are included and in 1976-77, 25 two-term offerings are included.

METHODOLOGY

The School of Business and Economics at Wilfrid Laurier University had used an evaluation form, at the instructor's option for some time, but in April, 1975, the Faculty Council approved a new form and a set of proposals regarding its administration. The form and rules governing its use were designed primarily for administrative purposes.

The evaluation in all courses is conducted in the tenth week of a thirteen week course, generally after mid-term marks are available but before major assignment or final grades are known. The instructor is informed of his (her) evaluation results only after final grades have been submitted and marks meetings have been held.

This paper reports on two aspects of student evaluations at Wilfrid Laurier University. First, the relationship between student evaluations and a number of independent variables is reported. This phase of the study covers the period 1975 to 1977 and was conducted during 1978. The second portion of the paper reports on faculty reaction to a basically administratively-oriented process. The latter phase was conducted in 1979 to determine whether student evaluation of teaching had any instructional impact on those being evaluated since that remains the ultimate goal of even an evaluation carried out for administrative purposes.

The School of Business and Economics teaches mainly undergraduates, with students being registered in two departments. About half the students' courses are taken within the School while about half are taken in options, normally given by the Faculty of Arts and Science. Table 1 details the registration in the terms being examined. Column 1 details the number of course registrants; column 2, the number of courses and sections offered; and column 3, the number of courses and sections for which full information was available including evaluations, mark distributions, and instructor information.

TABLE 2

Number of Classes by Percentage of 'A' Grades
By Instructor Evaluation

Evaluation of Instructor	Percentage of Students Receiving Grade of A			
	Less Than 15%	16% to 25%	26% to 40%	Greater Than 40%
Less than 4.5	27	11	11	3
4.5 - 4.9	17	22	7	0
5.0 - 5.4	29	21	9	6
5.5 - 5.9	25	23	25	9
6.0 - 6.4	13	16	21	10
Greater than 6.5	0	5	4	4

TABLE 3

Student Evaluation of Instructor By Time of
Day When Course Offered

Evaluation	Time Classes Offered			
	Early Morning	Mid-Day	Afternoon	Evening
0 - 4.4	18	6	20	8
4.5 - 4.9	13	9	20	3
5.0 - 5.4	25	20	14	6
5.5 - 5.9	36	17	25	3
6.0 - 6.4	13	16	25	6
6.5 - 7.0	8	2	3	0

The Chi-square is 27.22 which is significant at the .05 level.

Marks, as a percentage distribution were made available by the Registrar's Office. For most courses and sections, the marks distribution provided no problem. However, Wilfrid Laurier University does teach business courses to University of Waterloo co-op math and recreation students and these marks were submitted separately. When the number of University of Waterloo students constituted a large percentage, as they did in the Introductory Business Course, the grade distribution of Wilfrid Laurier students was

not considered sufficiently representative of the class so those classes were eliminated from the study.

Other data included the time of day the course section was offered, the enrolment, the year in which the course would normally be taken, whether the course was compulsory, whether the instructor had taught the course previously, the length of the class, whether there was a Friday class, and whether the class was a special section for University of Waterloo students.

FINDINGS

Clearly, the relationship between grades and instructor evaluations is of greatest interest. Here, a chi-square between the percent A's given and the overall evaluation of the instructor is significant at the .005 level which would initially lead one to assume that grades and evaluations of instructors are not randomly unrelated (chi-square = 43.76). This is seen in Table 2 where the classes with low evaluations of the instructor tend to have lower percentages of A grades awarded and vice versa.

However, when the relationship between the percentage of students attaining an 'A' grade and the evaluation of the instructor is controlled for the year in which the course is normally offered, the statistical significance of the relationship disappears. Similarly, when the relationship between grades and instructor evaluation is controlled for whether the course is compulsory or not and whether it is taught by full-time faculty or part-time faculty, the statistical significance of the relationship is decreased although for non-compulsory courses the relationship is still significant at the .005 level (chi-square = 33.84).[3]

There are other relationships to examine, as well as the relationship between grades and evaluations. These other relationships including time of day of classes, and size of class can be just as important as grades in determining the evaluation given by the student. For instance, the differences in evaluations between classes offered at various times of the day are significant at the .05 level (chi-square = 27.22).

In Table 3 it is noted that there is a statistically significant variation depending on the time of day. More early morning classes (starting before 10.30 a.m.) are rated at a higher level than are courses starting at other times during the day. No classes starting after 5.30 p.m. give instructors top evaluations.

While it is possible that instructors are fresher and more capable in the early morning than during the remainder of the day, that is an unlikely reason for the relaionship. It is probable that a self-selecting process is taking place. "Good" students are either too busy to arrange classes during other hours or they do not object to an early morning start. In either case, "good" students probably rate "good" instructors highly. Furthermore, weaker students who get into early morning classes are more likely to miss the evaluation (which takes place at the start of the class) leading to a higher evaluation score because only the highly motivated students are present.

The data collected for this study permits an examination of the influence on the evaluation of instructor by class size. The results are set out in Table 4 in which it can be seen that small classes get more of the highest evaluations than do large classes. Small classes also have more of the lowest evaluations than would be predicted but this is not as overpowering as the number of higher evaluations.

In small classes, the mean evaluation of the instructor can be heavily influenced by the student giving the highest score and the one giving the lowest score. Large classes do

TABLE 4

Student Evaluation of Instructor By Number of Evaluation

Evaluation	Less than 20 Evaluations	Between 20 - 39 Evaluations	Between 40 and 59 Evaluations	More than 60 Evaluations
0 - 4.4	13	33	2	4
4.5 - 4.9	10	27	8	1
5.0 - 5.4	23	33	9	0
5.5 - 5.9	22	38	22	0
6.0 - 6.4	21	17	18	4
6.5 - 7.0	9	4	0	0

The Chi-square is 53.00 which is significant at better than the .0005 level.

TABLE 5

Faculty Reaction to Student Evaluation

Question	No Response	Number Responding					Average Response
		Not at all 1	2	3	4	A great deal 5	
Examine computer print-outs	0	0	1	5	8	11	4.16
Adjusted course content	2	6	7	5	2	3	2.52
Adjusted textbook	2	11	3	4	2	3	2.26
Adjusted method of presentation	1	4	4	7	6	3	3.00
Adjusted grading to easier	3	19	1	1	1	0	1.27
Adjusted course organization	2	7	5	5	2	4	2.61
Adjusted explanation of course content	1	4	7	5	4	4	2.88
Adjusted availability	2	17	2	4	0	0	1.43
Adjusted time allotted to topics	2	11	5	3	3	1	2.04

not experience the same difficulties because one student cannot influence the results to such an extent.

One way to alleviate the problem of bias in small classes would be to eliminate the highest evaluation and the lowest evaluation in *all* classes. This would affect student ratings of instructors in small classes but not in large classes and then possibly the chi-square for size of class would not be so significant. Elimination of the highest and lowest evaluations from the numerical calculations is easily programmable. There are some people who would argue that to eliminate any evaluations is not morally correct. What is being posited here is not an elimination from the printout of raw scores but rather the elimination of the highest and lowest score in calculating the mean. The instructor who received one evaluation of 1, two evaluations of 2, three evaluations of 3, and one evaluation of 7 would receive a mean of 2.60 instead of 3.00 as would occur currently. In a large class of sixty students where one student gives a 1, 10 students a 4, 20 students a 5, 25 students a 6, 4students a 7, the newly calculated mean would be 5.36 instead of 5.32, hardly any change at all. While it is important to note that a small minority (even of one) strongly dislike a faculty member or strongly favour a faculty member, the purpose of the evaluation is to determine whether the faculty member is performing very well or very poorly as far as the majority of students are concerned.

INSTRUCTOR REACTIONS

Although it has been suggested that student evaluations are not really measures of productivity, they may influence the professor's inputs into the production process. To test this hypothesis, all faculty in the School of Business and Economics were surveyed in winter term 1979. Questions were asked about the influence student evaluations had on course content, textbook, presentation, grading, and organization. As well, faculty were asked to agree or disagree with a number of statements. Individual faculty members were asked to respond to questions about courses they had taught more than once during the period being examined. The response to the general questions about student evaluations worked out to be 50%, whereas the response to questions about individual courses was much smaller and most respondents who did reply (91%) indicated that they personally had no knowledge about why the evaluations were different. One faculty member indicated that the textbook was changed giving him (her) a "better" relationship with the class and another faculty member indicated that with a small class, one student had biased the results by a series of ones in one of the years (in this case the course evaluation was 4.3 in one year and 5.5 the next).

Tables 5, 6, and 7 examine the general reactions of faculty to student evaluations. In Table 5, it is evident that while the faculty do examine the computer print-outs of evaluations, they indicate little response to those evaluations. The method of presentation, including group work and audio-visuals was the most likely change to be made by the faculty member to the evaluations. Even here, only 37 percent of the faculty members indicate a great deal of adjustment or a large amount of adjustment. Faculty members are most emphatic in indicating that they do not change grading standards in response to student evaluations.

Table 6 indicates, though, that more than one-half the faculty members believe that if they give higher grades they will receive higher evaluations. Surely, there is an unconscious pressure to increase grades despite the stated profession otherwise. Other beliefs about

TABLE 6

Faculty View of Student Evaluation

	Statement	Percentage of Those Responding to Statement Agreeing
i)	If I give high grades, I get better evaluations	52.4
ii)	I am a reasonable teacher, so why should I bother trying to improve	12.0
iii)	Student evaluations can only pick out the superstars and the dregs	34.8
iv)	Administrators look at student evaluations very carefully	79.2
v)	There is nothing I can do about my personality so why should I worry about looking at the evaluations	4.0
vi)	I have talked to my chairman or dean about how to improve my student evaluations initiated by me	24.0
vii)	I believe that student evaluations are a popularity contest	45.5
viii)	I value the information I can derive from student evaluations	76.0
ix)	Ratings are insufficient, I need the written comments to determine problems	68.0
x)	I believe that student evaluations are a measure of teacher productivity	29.2
xi)	I believe that student evaluations have helped me improve my teaching	80.0
xii)	Student evaluations give me an ego boost	54.5
xiii)	I have received extra merit pay because of good student evaluations	22.7
xiv)	Merit is based on research and committee work and has nothing to do with student evaluations of teaching	9.1
xv)	I would just as soon drop evaluations but they do let students release some hostility at little cost	25.0
xvi)	I have talked to my chairman or dean about how to improve my student evaluations in a discussion initiated by the dean or chairman	8.3

student evaluations emerge in Table 6. Almost half the faculty members responding indicate that student evaluations are a popularity context and less than one-third believe they measure productivity. One-quarter would just as soon drop the evaluations except for the harmless venting of student feelings which they permit. Three quarters of the faculty responding indicate that the numerical print-outs are not sufficient – they need the written comments to determine what adjustments, if any, are to be made to their courses.

This latter point is somewhat contradicted by the results of another question asked. Faculty members were asked to rank the effectiveness of student input. The results are found in Table 7. There, the numerical evaluations are considered to be the most effective input to a faculty member's teaching while the written comments are, according to the faculty members, the least effective. Of course, what may be reflected here is the fact that an agreement between faculty and students currently precludes the distribution of written comment sheets until after grades have been submitted and that is further delayed if the course is sequenced with the same professor teaching another related course in the next term.

TABLE 7

Importance of Method of Feedback

Question	No Response	Number Responding Most important 4	3	2	Least important 1	Average Response
Numerical Evaluation	3	4	7	8	3	3.22
Written Comments	3	5	5	3	9	2.27
Class or Office Comments from students	3	2	7	10	3	2.36
Comments through Dean or Chairman	6	10	1	1	7	2.74

To summarize the results of the survey of faculty reactions to the student evaluations is made difficult by the widely divergent views held by the faculty members. It is clear, however, that faculty are not impressed enough by the evaluations to make changes to their teaching. Moreover, the belief that the evaluations do not measure productivity, but rather popularity is widespread. Thus, most faculty appear willing to support the continuation of the evaluations but to attribute that feeling to anything more than a potential ego boost would be wrong. Over ninety percent of the faculty do not believe the evaluations are responsible for extra merit pay.

If it can be argued that anything which causes a faculty member to pay attention to the teaching function is better than nothing and that most faculty are sincere in their effort to be conscientious teachers, student evaluations do serve the function of drawing attention to the teaching function.

CONCLUSIONS AND RECOMMENDATIONS

There appears to be no validity to the suggestion that a professor can influence his own evaluation by the grades he gives students. The percent of A's given in course controlled for year in which course is taught is not significantly related to the overall evaluation of the instructor. The fact that in fourth year it does vary can be explained by other factors.

Furthermore, it has been argued in the paper that certain other factors are more likely to influence student evaluations than grades. Early morning classes show a statistically significant higher evaluation as do small classes. Whether the course is compulsory, whether the course is taught by faculty with full-time appointments, and whether the faculty member is teaching the course for the first time, all generate differences in instructor evaluations.[4]

Obviously, if the purpose of students evaluating instructors is to facilitate the administrative award of merit or promotion, it would appear that there are certain factors which should be taken into account. Most importantly, the person trying to interpret evaluations of instructors should be careful to weight the results by the time of day the class started and by class size because these are factors over which the instructor has little individual control.

It would probably be worth experimenting with the suggestion that the top evaluation and the bottom evaluation be removed from the mean as suggested in this paper. This would help to decrease the differences between large and small classes. Furthermore, a person or department trying to interpret an instructor's evaluations should take into account whether the course is being taught for the first time by that instructor or whether the course is compulsory. If either is the case, the evaluation would be expected to be lower.

One very striking result, when faculty reaction is examined is the complete schizophrenia amongst the faculty. Contrary opinions and inconsistencies are common and the result is that the evaluations lose much of their potential as a vehicle for teaching improvement.

From the results of this paper, it would appear that student evaluations of instructors do in fact serve a purpose in relating information to administrators, students, other instructors, and most importantly to the instructor himself. However, there is some doubt about the usage of such information by instructors. Further, there is some question about whether the administrators should place heavy reliance on the evaluations. The instructor cannot strongly bias the results by easy grading but the results are subject to administrative influence in that timetabling, class size, and course assignment can all affect the evaluation. Interpersonal and intertemporal comparisons are made difficult because of this and if one wishes to stimulate faculty towards better teaching, merit rewards should be given only after account has been taken of these factors. The latter would presumably cause faculty to pay more attention to teaching and less attention to trying to influence the timetable or their own course load.

FOOTNOTES

1. The author acknowledges the financial assistance of an Instructional Development Grant from Wilfrid Laurier University as well as the research help of Michael Collins.
2. As reported in R.I. Miller (1974).
3. The breakdown of classes by year, by full-time or part-time faculty, and by compulsory vs. non-compulsory courses is available from the author.
4. The author will gladly supply the data.

REFERENCES

Anikeef, A.M. Factors Affecting Student Evaluation of College Faculty Members. *Journal of Applied Psychology*, 1953, *37*, 458-60.

Brown, D.L. Faculty Ratings and Student Grades: A University-wide Multiple Regression Analysis. *Journal of Educational Psychology*, 1976, *68*, 573-578.

Caffrey, B. Lack of Bias in Student Evaluation of Teachers. *Proceedings of the 77th Annual Convention of the American Psychological Association*, 1969, *4*, 641-2.

Costin, F., Greenough, W.T., & Mengers, R.J. Student Ratings of College Teaching: Reliability, Validity and Usefulness. *Review of Educational Research*, 1971, *41*, 511-35.

Feldman, K.A. Grades and College Students' Evaluations of their Courses and Teachers. *Research in Higher Education*, 1976, *4*, 69-111.

Jiobu, R.M., & Pollis, C.A. Student Evaluations of Courses and Instructors. *American Sociologist*, 1971, *6*, 317-321.

Miller, R.I. *Evaluating Faculty Performance.* San Francisco: Jossey-Bass Publishers, 1974.

Murray, H.G. *Evaluating University Teaching: A Review of Research.* Toronto: Ontario Confederation of University Faculty Associations, 1980.

Perry R.R., & Baumann, R.R. Criteria for the Evaluation of College Teaching: Their Reliability and Validity at the University of Toledo. In A.L. Sockloff (Ed.) *Proceedings of the First Invitational Conference on Faculty Effectiveness as Evaluated by Students.* Philadelphia, Pennsylvania; Measurement and Research Center, Temple University, 1973.

Pohlmann, J.T. A Multivariate Analysis of Selected Class Characteristics and Student Ratings of Instruction. *Multivariate Behavioural Research,* 1975, *10,* 81-91.

Rayder, N.F. College Student Ratings of Instructors. *Journal of Experimental Education,* 1968, *37,* 76-81.

Sherman, B., & Blackburn, R.T. Personal Characteristics and Teaching Effectiveness of College Faculty. *Journal of Education Psychology,* 1975, *67,* 124-131.

Voecks, V.W., & French, G.M. Are Student-Ratings of Teachers Affected by Grades? *Journal of Higher Education,* 1960, *31,* 330-334.