25/1/95

Potential Biases in Using Student Evaluations as Summative Teaching Performance Measures for Business Faculty

Geoffrey Poitras and George Blazenko*

Faculty of Business Administration
Simon Fraser University
Burnaby, B.C.
CANADA V5A 1S6

ABSTRACT

This paper provides empirical evidence on sources of potential bias in using student teaching evaluations to assess the teaching performance of business instructors. The sample dependent variables are teaching and course response items for 1600 courses given in a Canadian business faculty between 1986 and 1993. Courses surveyed cover all functional areas of business studies, such as Marketing, Accounting and Finance. Empirical evidence is provided for statistically significant differences in the dependent variables across: functional areas, class size, type of appointment and course level. This raises the potential for institutionalized bias when student teaching evaluations are used for assessing teaching performance of business faculty.

* The authors thank Irene Gordon, Eng Choo and Daniel Shapiro for helpful contributions as well as the Office of the Dean, SFU Faculty of Business Administration, for providing the data.

Potential Biases in Using Student Evaluations as Summative Teaching Performance Measures for Business Faculty

The study of student evaluations has generated thousands of papers produced over more than six decades.[1] Despite considerable scrutiny given to the validity of the information contained in student teaching evaluations (STE), there is relatively little analysis on how to use the information from student evaluations in assessing faculty teaching performance for administrative purposes. This is a significant problem for business faculties where comparisons are being made between instructors using different pedagogy and teaching material with considerable variation in difficulty. The comparison process is further complicated by the continuing debate on some significant empirical issues surrounding the validity of STE, e.g., Cashin and Downey (1992), Marsh (1991), Abrami (1989). Adequate resolution of debate over the interpretation and use of STE is confounded by the difficulty in identifying 'teaching effectiveness' (Sherman and Blackburn 1975, Shmanske 1988, Scriven 1989), an unobserved variable which student evaluations are supposed to measure. If teaching effectiveness is multidimensional as claimed by Marsh (1984, 1991) and others, then a number of items need to be considered in order to assess teaching effectiveness. Because administrative efficiency typically dictates consideration of only one or two global response items in order to expedite decision making, there is a distinct possibility of bias in the use of STE as a summative measure for use in personnel decisions.

This paper examines student teaching evaluation responses to two commonly used summative measures: instructor teaching ability; and, course rating. The sample contains information on 1600 courses over 18 trimesters for the Faculty of Business Administration at Simon Fraser University. The Faculty is composed of seven **functional groupings**: Accounting, Marketing, Finance, Human Resources Management, Business Policy, Management Science/MIS, and International Business. Evidence is provided on the variation in student responses to teaching and course measures arising from: class size, level of course, type of instructor appointment and functional grouping. Some of these variables, such as course level and class size, have been exhaustively examined in previous studies, for both business and non-business school samples. However, there is almost no evidence on other variables, such as the variation in summative student evaluation responses across functional

groupings. Evidence of potential biases associated with these two global response items raises significant questions about the use of these measures for purposes of summative evaluation in important administrative decisions, such as tenure and promotion.

I. Previous Research

Cashin and Downey (1992), among others, recognize that interpreting the information in student evaluation responses depends fundamentally on whether formative or summative evaluation is involved. More precisely, **formative evaluation** is aimed at diagnosing and improving course and instructor performance. This will typically involve considering results from student evaluation surveys, covering a range of factors identified as important for teaching effectiveness. Individual response items may be considered, even though such measures are not highly correlated with specific objectives. In contrast, **summative evaluation** is useful for making personnel decisions, e.g., Thorne (1980), Miller (1987). In this case, administrative efficiency favours the use of summary measures of instructor and course performance. For such a procedure to be fair and equitable, identification of potential biases associated with summary measures is required, e.g., Sheehan (1975). If the bias is systematic, appropriate adjustments could be made to, at least, make the summative measures a better indicator of teaching effectiveness.

Despite the considerable amount of research on student evaluations, there is still on-going debate on various issues. In a comprehensive survey of the available empirical research on a range of response items contained in student evaluations, Marsh (1984) finds:

> …class average student ratings are: (a) multidimensional; (b) reliable and stable; (c) primarily a function of the instructor who teaches a course rather than the course that is taught; (d) relatively valid against a variety of indicators of effective teaching; (e) relatively unaffected by a variety of variables hypothesized as potential biases; and, (f) seen to be useful by faculty as feedback about their teaching, by students for use in course selection, and by administrators for use in personnel decisions. (p.707)

A number of these findings are considered debatable. The most contentious are: (a) multidimensionality, (d) validity and (e) unbiasedness. For example, validity requires specification of criteria for identifying and measuring teaching effectiveness, a decidedly difficult task. Consider the use of student learning as a potential criteria. This assumes that students of more effective instructors will learn more.[2] However, this approach requires the

measurement of another unobserved variable: student learning. Using grades in subsequent courses as a proxy for student learning, Watts and Bosshardt (1991) find evidence that student evaluations are **not** valid indicators of effective teaching. Using somewhat different criteria, Rodin and Rodin (1972), Dowell and Neal (1982), Shmanske (1988), and Gramlich and Greenlee (1993) find similar results.

Another key conclusion of Marsh (1984, 1991) is that teaching effectiveness is multidimensional. This result requires that a range of variables, such as course difficulty, instructor enthusiasm, subject knowledge and organization, interact in producing specific teaching outcomes. The practical implication is that a number of different items must be surveyed and considered in evaluating teaching effectiveness. In the absence of a high degree of correlation across response items, multidimensionality conflicts with the summative use of teaching evaluations. In opposition to multidimensionality, a number of researchers argue that teaching effectiveness can be largely measured with the use of a single global response item, typically identified with student responses to instructor effectiveness measures, e.g., Abrami (1989), Abrami and d'Apollonia (1991). In effect, there is one dominant **factor** in teaching effectiveness which has a high degree of correlation with the various multidimensional responses.[3] Again using student learning as a measure of teaching effectiveness, studies of whether specific instructors produced consistently abnormal student learning have produced conflicting results, with both weak and strong relationships for selected global measures being reported.[4]

The final contentious issue concerns the unbiasedness of student teaching evaluations, particularly the global response items.[5] At least partly because summative evaluation favours the use of global measures, considerable research has been dedicated to evaluating biases in specific global measures under a variety of sampling situations. Some of the variables which have been investigated include: class size; level of the course; expected or observed grade distributions; gender; and, instructor rank.[6] Conflicting results abound. For example, while Crittenden, et.al. (1975), Scott (1977) and others find large (small) class sizes tend to receive lower (higher) student evaluations, Aleamoni and Graham (1974), Jiobu and Pollis (1971) and others find no relationship. Some researchers, Marsh, et.al. (1979), Pohlmann (1975), report

a curvilinear relationship. These potentially conflicting results can partially be explained by substantive variation in relevant samples and research methodologies. However, even allowing for this source of variation, there are still potentially substantive sources of bias which could impact student evaluation responses.

While there are numerous studies of student evaluations for a variety of different teaching environments, studies involving samples from business schools are limited in number. Studies that are available typically examine a specific functional group, e.g., Marketing or Accounting, and are aimed at replicating results achieved using a non-business faculty sample. For example, Clayson and Haley (1990) use a sample of marketing courses to verify the Sherman and Blackburn (1975) result that instructor personality is the strongest factor influencing student evaluations. This result suggests that student responses to items such as "instructor accessibility" have more to do with personality than with actual office hours.[7] In another study, Hinkin (1991) examines a sample of undergraduate business school students to verify the Feldman (1978) result that the time of the day a course is offered has little impact on student evaluations. A final example, Scherr and Scherr (1990), examines a sample of finance courses for a number of sources of potential bias and finds evidence only for a bias with required/elective courses, as in McKeachie (1979).

While it is useful to have empirical evidence extending general results to business school samples, there are a number issues which are particular to business schools for which little or no evidence is currently available. For example, because business schools involve a multidisciplinary, integrated curriculum, it is often the case that administrators, e.g., faculty tenure committees, must evaluate the teaching performance of instructors from decidedly different functional areas. As a result, the potential for bias in summary measures across functional groupings has particular significance for business school samples. The limited evidence that is available, Arnett, et.al. (1989), finds a small bias across functional groups. Indirect evidence on this point, Tong and Bures (1987), finds that faculty in the two most rigourous functional groups, Finance and Management Science, perceive the instructor rating process to be unfair. Another issue of interest about which little empirical evidence is available is the widespread use of sessionals in business schools, a practice dictated by the

applied nature of business studies. On this issue, Goldberg and Callahan (1991) find evidence that including sessional instructors in the sample can introduce significant bias.

II. Sample Characteristics and Statistical Considerations

The sample consists of information from student evaluations on 1600 graduate and undergraduate courses given over 18 trimesters, Fall 1986 to Fall 1993, for the Faculty of Business Administration at Simon Fraser University (FBA).[8] Both executive and regular MBA courses are included. The actual student evaluations contain twenty-two response items, covering the range of conventional topics: student characteristics, difficulty/workload, grading, and instructor characteristics. From all possible response items, the summative evaluation process used by the FBA highlights student responses to the two global questions: 1) How do you rate the instructor's teaching ability? (Excellent = 4 to Poor = 0); and, 2) How do you rate this course? (Excellent = 4 to Poor = 0).[9] Responses are discrete, with students only allowed to assess an integer value between 0 to 4, inclusive. Additional information was also obtained for: functional area of the instructor, class size, instructor rank (regular vs. sessional), course, individual instructor, and the term the course was given. In the FBA, student evaluations are typically conducted in the period one to three weeks prior to the commencement of final exams. Because individual instructors in the FBA have substantial autonomy over course content even in required courses and in courses with multiple sections, the interpretation of the course content measure may differ from responses at institutions where course content is mandated externally.[10]

Determining the extent to which student evaluations of instructor performance are biased by specific extraneous factors is a difficult task, at best.[11] Various statistical techniques, including factor analysis, multiple regression and classical analysis of variance, have been used to examine different facets of the teaching evaluation problem. The statistical technique selected is typically motivated by the type of null hypothesis being tested. In the present case, the null hypotheses take the form of questions such as: do the mean teaching and course scores differ across groups? And, are the mean teaching and course scores stable across time? These types of questions are well-suited to testing by classical analysis of variance.[12] Within this

framework, tests for equality of means can be made using pairwise multiple comparisons to test hypotheses such as: which functional group means are significantly different from each other? For this purpose, Tukey's HSD test (e.g., Daniel 1978) has been adapted to specify the appropriate critical values for pairwise comparison of specific means.

There are a number of limitations in using the classical analysis of variance approach. One limitation is the difficulty of identifying multivariate interactions which contribute to total variance. For this purpose, multiple regression is a potentially more desirable statistical procedure. However, to be used correctly, multiple regression requires the formulation of specific null hypotheses. For present purposes, regression results would only be provide a data summary, without any specific reference to the underlying analytical model which determined the variable selection procedure. In addition, the discrete or polytomous nature of the many of the variables involved in the empirical analysis, such as course level and functional area, raise substantive questions about the properties of the resulting estimators. With these caveats, some regression evidence is also presented.

Another limitation of analysis of variance, and classical statistics in general, is the difficulty of choosing an appropriate critical value for hypothesis testing. When results from significantly different samples sizes are being compared, this problem is often manifested in a form referred to as Lindley's paradox which is associated with changes in the unjust acceptance (β) power of hypothesis tests as sample size increases, when holding constant the critical level for rejection power (α). This creates difficulties for the typical convention which involves making statements about statistical 'significance' using a fixed α for the test. Because in what follows, $\alpha = 5\%$ will be used to assess significance, this requires some caution to be used when comparing results from the F tests, which involve larger sample sizes, with tests involving smaller samples, such as the multiple comparison tests. (In the Tables, significance is indicated by a (*) following the reported value.)

III. Empirical Results

In the analysis of variance, a direct test of a specific hypothesis involves examining the ratio of the unexplained to the explained sum of squares attributable to a selected variable.

It is also possible to indirectly test hypotheses by censoring the sample thereby creating a nested null hypothesis. In practical terms, this is typically accomplished by providing two sets of results, for example using all faculty instructors for one sample and all faculty omitting sessional, nontenure track instructors for another sample.[13] Selection of the most appropriate sample censoring variables can often be motivated by institutional considerations, e.g., sessional instructors are usually given lower level courses to teach which tend to give lower-on-average teaching and course scores. Censoring of the sample can also be done using other variables, such as eliminating courses which are identified as being particularly difficult, so-called 'killer courses', from tests for a course level bias. Using this approach to ANOVA requires careful selection of the appropriate censoring variables in order to keep the number of reported results manageable.

A) Functional Groups and Instructor Rank

Tables 1 and 2 present distributional and ANOVA results for the faculty and each of the functional groups. Table 1 provides results for all courses given in the FBA over the sample with Table 2 providing results for a sample which is censored by **excluding the results from courses taught by sessional instructors**. A number of interesting conclusions emerge from these results. For example, consider the observed differences between the average values of the two global response items. The t-tests indicate that there is a significant difference between the faculty course and teaching scores. While the results are not reported, there are also significant t-tests for the difference between course and teaching scores for each of the functional groups. In all cases the teaching score is greater than the course score. (A similar comment can also be made about the **volatility** of the two measures, with the teaching score exhibiting the higher volatility.

From the perspective of using global measures for summative evaluation, this significant difference is of interest because of the substantively greater institutional emphasis placed on the teaching score in making administrative decisions in the SFU FBA. Examination of the Spearman rank correlation coefficients reveals that rankings based on the two summative measures will be somewhat different. The similarity of the Spearman and variate correlations reveals that the difference in rankings is likely not due to the presence of one or two extreme

observations. A similar conclusion can be drawn from the skewness and kurtosis results which indicate that the distributions for the two measures are typically close to normal. The difference between the teaching and course score responses is even more perplexing when it is considered that institutional features of the sample often make the instructor largely responsible for course content, implying that the observed difference is inconsistent with **a priori** expectations of no systematic difference.

There are a number of plausible explanations for the significant differences between course and teaching scores. One explanation is that the teaching score is recognized as the most important summative measure (see n.9) and may be subject to a feedback effect, where the importance of the teaching score as a summative measure leads to biases being introduced which affect student responses to this item. Stratton, et.al. (1994), for example, find that the student evaluation process had a significant impact on faculty grading procedures. Feedback would be consistent with a more general explanation proposed by Engdahl (1993) and others associated with the predictions of expectancy and attribution theory: "Students will logically take revenge on a person, the instructor; but it is illogical to take revenge on a course." In the current sample, the observed difference translates into **positive** effects on the teaching score originating from factors such as instructor personality and grading practices which do not have as large an impact on the students' evaluation of course content.[14]

Another important result in Tables 1 and 2 is the significant difference between the functional group means, for both the teaching and course scores. This effect was more significant when courses taught by sessionals were included in the sample. This is consistent with the institutional observation that sessionals tend to teach large, lower level classes, which tend to generate lower teaching and course scores. However, even when sessionals are excluded from the sample, there is still a significant difference in the functional group means. Given this, pairwise comparison tests can be used to provide further information on whether a specific pair of means is different. Taking the $\alpha = 5\%$ critical difference to be approximately $\pm .11$, when sessionals are included in the sample (Table 1), only the management science group mean for course rating is significantly different from the other group means. A similar result applies to the mean teaching score for the Finance group in

Table 2, with sessionals excluded. Because Management Science and Finance courses tend to have a more mathematical content, this result is consistent with course rigour producing a reduction in summative student evaluation scores.

B) Level of Course and 'Killer' Courses

Numerous studies for non-business school samples have concluded that the level of the course is not an important source of bias in student teaching evaluations. However, other studies have demonstrated significant bias between required and elective courses. In the FBA business curriculum, the lower the level of a course, the more likely that course will be required. Even in upper division courses which are required, students have already self-selected their functional area and would be likely to take the course if it were elective. In the present sample, the course level effect is also compounded with a lack of homogeneity in the student samples. For example, the 200-level courses, which are not taught by all functional groups, are prerequisites to FBA admission and are open to non-business students. The 300-level courses are a mix of FBA core courses, required of all FBA students, and area core requirements which will typically be taken only by area concentrators. The 400-level courses are taken primarily by area concentrators.

Due to factors such as the elective/required course bias, class size effects, different grading requirements, lack of sample homogeneity and so on, there is considerable **a priori** potential for bias to be associated with the level of the course. Table 3 confirms that the level of the course is a significant source of bias in both the teaching and course responses, with the score received generally increasing with the course level, with the exception of the 500 and 600 level graduate courses being lower than 400 level undergraduate courses. This result is likely due to the differences in the student samples, with the 600 Executive MBA level representing a student sample which is decidedly different from the others. By far, the best course and teaching scores were recorded in the 800 level advanced graduate classes. As in Tables 1 and 2, there is again a significant difference between teaching and course scores, for all course levels.

A number of tests were conducted to investigate whether the course level results were due to the impact of certain 'killer' courses which typically received scores significantly below the

faculty average. Six such courses were identified. Each of these courses was lower (200 and 300) level and emphasized quantitative capabilities. As indicated in Table 4, when the teaching and course score results for these courses was removed from the sample, the course level effect is found to be less significant. The degree of bias for specific courses is illustrated in Table 5, where results for the two lowest rated courses is reported. It can be observed that the **maximum** course score received in these killer courses over the sample is not substantially above the **average** score for the advanced 800 level graduate courses. Another interesting result occurs for FIN 312 where, despite near equality in the average course and teaching scores, there is still a significant difference in volatility.

C) Other Factors: Temporal Stability and Class Size

Previous research on student teaching evaluations has confirmed the stability of results across a variety of sampling situations. For example, Aleamoni (1987) reports correlations over time of 0.87 to 0.89. For the full sample, including all courses and types of instructors, the following analysis of variance results were obtained:

$F(6, 1593)$ test of equality of Course Means Across Years: 0.89
$F(6, 1593)$ test of equality of Teaching Means Across Years: 2.67*

The significance of the differences in the teaching means across time presents a potentially anomalous result. However, when sessional instructors are excluded from the sample:

$F(6, 1019)$ test of equality of Course Means Across Years: 0.40
$F(6, 1019)$ test of equality of Teaching Means Across Years: 0.849

As with the results for bias across functional groups, there is some evidence that sessional instructors introduce a systematic, albeit small, source of bias. This result is due to the variation in the sessional instructor pool across time and, possibly, to the introduction of certain 200 level courses (taught primarily by sessionals) during the sample. Exclusion of sessional instructors is sufficient to produce the expected result: the average responses for both the course and teaching items are stable from one year to the next.

Regarding the impact of class size, results already presented provide a strong indication that this variable will be highly significant. For example, the bulk of classes at lower levels are larger than more advanced classes. In the present sample, only four classes at the 400

level and above are larger than 40 students. As a consequence, the significant bias associated with course level is likely to be compounded with a class size effect. This hypothesis is confirmed by examining ANOVA results for significant bias when classes are classified as: larger than 40; between 10 and 40; and, smaller than 10:

$F(2, 1597)$ test of equality of Course Means Across Class Size: 76.8*
$F(2, 1597)$ test of equality of Teaching Means Across Class Size: 149.6*

In order to control for the impact of course level on the ANOVA, a sample of only 400 level courses was examined. Four groups of class sizes were considered: less than 10; between 10 and 20; between 20 and 30; and greater than 30. The ANOVA results in this case are:

$F(3, 1596)$ test of equality of Course Means Across Class Size: 9.15*
$F(3, 1596)$ test of equality of Teaching Means Across Class Size: 5.40*

Multiple comparison tests reveal that of the six possible differences only the differences involving the small class mean is different. More precisely, small classes were found to produce significantly higher course and teaching scores.

D) Multiple Regression Results

One difficulty with classical analysis of variance is that relationships between the dependent and independent variables are often multivariate and complicated. These types of estimation problems often lead to the use of multiple regression as a method of summarizing data relationships. However, due to problems such as multicollinearity and omission of variables, regression analysis also presents certain difficulties in interpreting results. Because in analysis of variance results have already been obtained, multiple regression is being used to provides a summary of the ANOVA results. The regression results presented in Tables 6 and 7 provide evidence of significant bias associated with class size, course level, and type of instructor. The significant coefficient for year in the teaching regression confirms the result, for the full faculty sample, given in Section C. The regression results for functional areas indicates that the bias is not spread evenly across functional groupings. For example, the course score regression for Marketing indicates no significant sources of bias, while Management Science has a highly significant course level coefficient. Comparison of the R^2 across groups also reveals a somewhat different intergroup pattern of bias between the

teaching and course score results.

IV. Conclusions

The multidisciplinary nature of business school faculty raises a number of substantive questions about the use of summative teaching measures for administrative purposes, such as tenure and promotion decisions. In addition to differences in course content across functional groups, there is also significant variation across teaching methods, e.g., cases versus lectures, and the rigour of instruction. On balance, there is considerable potential for biases to emerge when global response items from student evaluations of instructors are used as summative measures for faculty personnel decisions. This paper considered whether such biases occur across: functional groups, class size, course level, and type of instructor. Using classical analysis of variance techniques, each of these factors was found to generate statistically significant bias in global teaching and course response items. In addition, for almost all situations considered, the average teaching score was found to be significantly higher, and more variable, than the course score. While it is possible that some or all of these sources of bias may arise due to institutional features associated with the sample, at least some of these types of bias will almost certainly appear in other samples.

Bibliography

Abrami, P., "How Should we use Student Ratings to Evaluate Teaching?" Research in Higher Education (1989): 221-7.

Abrami, P. and d'Apollonia, "Multidimensional Students' Evaluations of Teaching Effectiveness-Generalizability of 'N= 1 ' Research", Journal of Educational Psychology (1991): 411-15.

Aleamoni, L., "Student Rating Myths Versus Research Facts", Journal of Personnel Evaluation in Education (1990): 111-19.

Aleamoni, L. and M. Graham, "The Relationship between CEQ Ratings and Instructor's Rank, Class Size and Course Level", Journal of Educational Measurement (1974): 189-202.

Aleamoni, L. and P. Hexner, "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 (1980): 67-84.

Arnett, K., D. Arnold, and D. Cochran, "Improving Business School Student Evaluation of Faculty Performance", Journal of Education for Business, (March 1989): 268-70.

Baird, J., "Perceived Learning in Relation to Student Evaluation of University Instruction", Journal of Educational Psychology (1987): 90-91.

Cashin, W. and R. Downey, "Using Global Student Rating Items for Summative Evaluation", Journal of Educational Psychology (1992): 563-72.

Clayson, D. and D. Haley, "Student Evaluations in Marketing: What is Actually Being Measured?", Journal of Marketing Education: (Fall 1990).

Crittenden, K, J. Norr and R. LeBailly, "Size of University Classes and Student Evaluations of Teaching", Journal of Higher Education (1975): 461-70.

Daniel, W., Biostatistics (2nd.ed.), New York: Wiley (1978).

Dowell, D. and A.J. Neal, "A Selective Review of the Validity of Student Ratings of Teaching", Journal of Higher Education (1982): 51-62.

Engdahl, R., R. Keating and J. Perrachione, "Effects of Grade Feedback on Student Evaluation of Instruction", Journal of Management Education (1993): 174-84.

Etherington, L., "Toward a Model of Accounting Pedagogy: A Critical Incident Analysis", Issues in Accounting Education (Fall 1989): 309-26.

Feldman, K., "The Superior College Teacher from the Student's View", Research in Higher Education (1976): 243-88.

_____, "Course Characteristics and College Students; Ratings of their Teachers: What we know and what we don't", Research in Higher Education (1978): 199-242.

Goldberg, G. and J. Callahan, "Objectivity of Student Evaluations of Instructors", Journal of Education for Business (July 1991): 377-8.

Gramlich, E. and G. Greenlee, "Measuring Teaching Performance", Journal of Economic Education (Winter 1993): 3-13.

Hinkin, T., "The Effects of Time of Day on Student Teaching Evaluations: Perceptions versus

Reality", Journal of Management Education (1991).

Jiobu, R. and C. Pollis, "Student Evaluations of Courses and Instructors", The American Sociologist (1971): 317-21.

Kemp, B. and G. Kumar, "Student Evaluations: Are We Using Them Correctly", Journal of Education for Business (Nov. 1990): 106-11.

Marlin, J. and J. Niss, "End-of-Course Evaluations as Indicators of Student Learning and Instructor Effectiveness", Journal of Economic Education: (Spring 1980): 16-27.

Marsh, H., "Multidimensional Students' Evaluations of Teaching Effectiveness: A Test of Alternative Higher-Order Structures", Journal of Educational Psychology (1991): 285-96.

_____, "Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research", International Journal of Educational Research (1987): 253-388.

_____, "Student Evaluations of University Teaching: Dimensionality, Reliability Validity, Potential Biases, and Utility", Journal of Educational Psychology (1984): 707-54.

Marsh, H., J. Overall and S. Kesler, "Class size, student evaluations and instructional effectiveness", American Educational Research Journal (1979): 57-69.

McKeachie, W., "Student ratings of faculty: a reprise", Academe (1979): 384-97.

Miller, R., Evaluating Faculty for Promotion and Tenure, San Francisco: Jossey-Bass (1987).

Pohlmann, J., "A multivariate analysis of selected class characteristics and student ratings of instruction", Multivariate Behavioral Research (1975): 81-91.

Remmers, H., "The Relationship Between Students' Marks and Students' Attitudes Towards Instructors", School and Society (1928): 759-60.

_____, "To What Extent do Grades Influence Student Ratings of Instructors?" Journal of Educational Research (1930): 314-6.

Rodin, M. and B. Rodin, "Student Evaluations of Teachers", Science (1972): 1164-66.

Scherr, F. and S. Scherr, "Bias in Student Evaluations of Teacher Effectiveness", Journal of Education for Business (May 1990): 356-58.

Scott, C., "Student Ratings and Instructor-Defined Extenuating Circumstances", Journal of Educational Psychology (1977): 744-47.

Scriven, M., "The Design and Use of Forms for the Student Evaluation of Teaching", Instructional Evaluation (1989): 1-13.

Sheehan, D., "On the invalidity of student ratings for administrative personnel decisions", Journal of Higher Education (1975): 687-700.

Sherman, B. and R. Blackburn, "Personal Characteristics and Teaching Effectiveness of College Faculty", Journal of Educational Psychology (1975): 124-31.

Shmanske, S., "On the Measurement of Teaching Effectiveness", Journal of Economic Education (Fall 1988): 307-14.

Stratton, R., S. Myers, and R. King, "Faculty Behavior, Grades and Student Evaluations",

Journal of Economic Education (Winter 1994): 5-15.

Tanner, J., H. Manakyan and D. Hotard, "Management-Faculty Research Productivity and Perceived Teaching Effectiveness", Journal of Education for Business (May 1992): 261-65.

Thorne, G., "Student Ratings of Instructors, From Scores to Administrative Decisions", Journal of Higher Education (1980): 207-214.

Tong, H. and A. Bures, "An Empirical Study of Faculty Evaluation Systems: Business Faculty Perceptions", Journal of Education for Business (April 1987): 319-21.

Watts, M. and W. Bosshardt, "How Instructors Make a Difference: Panel Data Estimates from Principles of Economics Courses", Review of Economics and Statistics (1991): 336-40.

Table 1

Summary Information on Teaching and Course Items
by Faculty and Functional Groups[#]

Sample: NOB = 1600 (Includes All Courses and Sessionals)

Item	Mean	Std.D.	Min.	Max.	Skew	Kurtosis
			FACULTY			
Course	2.97	0.408	0.80	4.00	-0.192	0.484
Teaching	3.15	0.519	0.20	4.00	-0.723	0.693

t-value for difference in Means for Course and Teaching: 10.91*

Spearman Rank Correlation: 0.778 Variate Correlation: 0.782

			Marketing NOB = 268			
Course	3.00	0.390	0.80	4.00	-1.13	4.21
Teaching	3.11	0.51	0.20	4.00	-1.50	4.71
			Human Resources Management NOB = 242			
Course	3.05	0.400	1.85	3.88	-0.24	-0.24
Teaching	3.22	0.481	2.00	4.00	-0.76	-0.08
			Business Policy NOB = 227			
Course	3.03	0.496	1.65	4.00	-0.30	-0.58
Teaching	3.25	0.585	1.25	4.00	-0.77	-0.05
			Management Science NOB = 263			
Course	2.83	0.389	1.64	4.00	0.24	0.31
Teaching	3.08	0.496	1.29	4.00	-0.37	0.16
			Finance NOB = 146			
Course	2.98	0.394	1.64	3.90	-0.25	0.72
Teaching	3.07	0.464	1.71	4.00	-0.38	0.01
			International Business NOB = 14			
Course	3.06	0.526	2.06	3.64	-1.04	-0.04
Teaching	3.24	0.504	2.33	3.86	-0.91	-0.27
			Accounting NOB = 440			
Course	2.95	0.360	1.74	4.00	0.22	0.50
Teaching	3.14	0.523	1.59	4.00	-0.61	-0.23

ANOVA F(6,1593) test for equality of Group Course Means: 8.49*
ANOVA F(6,1593) test for equality of Group Teaching Means: 4.15*

[#] NOB is Number of Observations. The two-tailed t test is for equality of Faculty Course and Teaching Scores; Kurt is the standardized fourth moment for kurtosis, centered about 3, the value for the normal distribution; Skew is the standardized third moment for skewness. (*) indicates statistical significance at the 5% level.

Table 2

Summary Information on Teaching and Course Items
by Faculty and Functional Groups[#]

Sample: NOB = 1026 (Includes All Courses and **Excludes** Sessionals)

Item	**Mean**	**Std.D.**	**Min.**	**Max.**	**Skew**	**Kurtosis**
			FACULTY			
Course	3.03	0.400	1.64	4.00	-0.182	0.161
Teaching	3.22	0.501	1.18	4.00	-0.760	0.523

t-value for Difference in Means for Course and Teaching: 12.2*
Spearman Rank Correlation: 0.813 Variate Correlation: 0.812

			Marketing NOB = 134			
Course	3.03	0.365	1.81	4.00	-0.45	0.91
Teaching	3.23	0.488	1.18	4.00	-1.10	2.07
			Human Resources Management NOB = 185			
Course	3.06	0.402	1.85	3.88	-0.12	-0.39
Teaching	3.22	0.470	2.00	4.00	-0.65	-0.16
			Business Policy NOB = 161			
Course	3.11	0.459	1.65	4.00	-0.63	0.40
Teaching	3.25	0.579	1.25	4.00	-0.81	0.17
			Management Science NOB = 163			
Course	2.94	0.379	2.08	4.00	0.30	-0.07
Teaching	3.18	0.495	1.29	4.00	-0.91	1.28
			Finance NOB = 140			
Course	2.98	0.383	1.64	3.87	-0.41	0.81
Teaching	3.08	0.464	1.71	4.00	-0.39	0.06
			International Business NOB = 11			
Course	3.19	0.329	2.52	3.60	-1.13	1.02
Teaching	3.35	0.339	2.54	3.78	-1.33	2.61
			Accounting NOB = 232			
Course	3.06	0.390	2.00	4.00	0.04	0.32
Teaching	3.27	0.496	1.59	4.00	-0.82	0.58

ANOVA F(6,1019) test for equality of Group Course Means: 3.57*
ANOVA F(6,1019) test for equality of Group Teaching Means: 2.65*

[#] t test is for equality of Faculty Course and Teaching Scores; Kurt is the standardized fourth moment for kurtosis, centered about 3, the value for the normal distribution; Skew is the standardized third moment for skewness. See also Notes to Table 1.

Table 3

Summary Information on Teaching and Course Items
by Course Level[#]

Sample: NOB = 1600 (Includes All Courses and Sessionals)

Item	Mean	Std.D.	Min.	Max.	Skew	Kurtosis
		200 Level NOB = 239				
Course	2.75	0.326	1.64	4.00	0.26	0.72
Teaching	2.98	0.534	1.62	4.00	-0.20	-0.81
		300 Level NOB = 522				
Course	2.87	0.336	1.65	3.75	-0.34	0.26
Teaching	3.05	0.485	1.47	4.00	-0.35	-0.39
		400 Level NOB = 552				
Course	3.06	0.389	1.57	4.00	-0.32	0.66
Teaching	3.25	0.461	1.13	4.00	-0.97	1.98
		500 Level (First Year DMBA) NOB = 25[##]				
Course	3.01	0.408	2.07	3.64	-0.72	0.18
Teaching	3.20	0.672	1.29	3.93	-1.34	1.53
		600 Level (Executive MBA) NOB = 120				
Course	2.99	0.53	0.80	3.90	-1.03	1.52
Teaching	3.01	0.65	0.20	4.00	-1.20	1.88
		800 Level (Second Year DMBA) NOB = 147[##]				
Course	3.32	0.391	2.20	4.00	-0.59	0.08
Teaching	3.52	0.388	2.20	4.00	-0.88	0.60

ANOVA F(5,1594) for equality of Course Means across Levels: 55.05*
ANOVA F(5,1594) for equality of Teaching Means across Levels: 32.70*

[#] Kurt is the standardized fourth moment for kurtosis, centered about 3, the value for the normal distribution; Skew is the standardized third moment for skewness. See n.8 for a further explanation of course numbering.

[##] Differences in the number of observations between the first and second year Day (full time) MBA levels is due to a change in the structure of the DMBA during the sample period. Prior to 1990, the DMBA was a specialist program which consisted only of 800 level courses, for which an undergraduate business degree was required for admission. After 1990, a second module was added to the program which admitted non-business undergraduate degree students who were required to take a sequence of required 500 Level, introductory business courses.

Table 4

Summary Information on Teaching and Course Items
by Course Level (Including Sessionals) #

Sample: NOB = 1387 (All Courses Except **Six** Killer Courses)

Item	**Mean**	**Std.D.**	**Min.**	**Max.**	**Skew**	**Kurtosis**
			FACULTY			
Course	3.02	0.398	0.80	4.00	-0.310	0.843
Teaching	3.19	0.506	0.20	4.00	-0.865	1.216
			200 Level NOB = 118			
Course	2.89	0.318	2.17	4.00	0.37	0.32
Teaching	3.12	0.529	1.71	4.00	-0.51	-0.85
			300 Level NOB = 430			
Course	2.92	0.320	1.74	3.75	-0.35	-0.40
Teaching	3.06	0.467	1.71	4.00	-0.42	-0.34

ANOVA F(4,1383) test for equality of Means for Course Level: 22.79*
ANOVA F(4,1383) test for equality of Group Teaching Means: 12.34*

* Three 200 and three 300 level courses were excluded. Because this means that the results for 400-800 are unchanged from Table 3, these results have not be duplicated in this Table. Kurt is the standardized fourth moment for kurtosis, centered about 3, the value for the normal distribution; Skew is the standardized third moment for skewness.

Table 5

Summary Information on the **Two** Lowest Rated Killer Courses
(Including Sessionals)

Item	**Mean**	**Std.D.**	**Min.**	**Max.**	**Skew**	**Kurtosis**
		BUS 232: Introductory Statistics NOB = 41				
Course	2.59	0.330	1.64	3.25	-0.45	0.69
Teaching	2.95	0.590	1.64	3.86	0.06	-0.82
		BUS 312: Introductory Corporate Finance NOB = 24				
Course	2.75	0.223	2.23	3.25	0.07	0.51
Teaching	2.77	0.407	2.12	3.59	0.10	-0.80

* Kurt is the standardized fourth moment for kurtosis, centered about 3, the value for the normal distribution; Skew is the standardized third moment for skewness.

Table 6

Multiple Regression Results for Course Responses*

Sample: NOB = 1600 (Includes All Courses and Sessionals)

Dependent Variable: Course Score (C)

Estimated Equation: $C = a_0 + a_1$ Class Size $+ a_2$ Course Level $+ a_3$ Area $+ a_4$ Year $+ a_5$ Graduate Dummy $+ a_6$ Sessional Dummy

a_0	a_1	a_2	a_3	a_4	a_5	a_6	R^2
			FACULTY				
2.41	-0.002	.002	.0008	.0003	-.05	.069	.17
(5.99)	(6.16)	(0.51)	(8.67)	(0.75)	(3.56)	(3.29)	
			Marketing NOB = 268				
1.85	-0.010	.0001	.0012	.0315	.04	.02	
(1.05)	(0.74)	(0.42)	(0.99)	(0.78)	(0.70)		
			Human Resources Management NOB = 242				
1.29	-0.005	.0002	.0021	.0032	-.11	.16	
(1.23)	(4.78)	(0.92)	(1.78)	(0.09)	(1.67)		
			Business Policy NOB = 227				
1.28	-0.009	.0010	.0017	-.0897	.14	.32	
(1.02)	(5.56)	(3.57)	(1.23)	(2.11)	(2.05)		
			Management Science NOB = 263				
2.18	-0.001	.0013	.0001	-.0367	.08	.37	
(2.56)	(1.42)	(8.08)	(0.09)	(1.36)	(1.84)		
			Finance NOB = 146				
4.26	-0.003	.0010	.0015	-.0994	-.14	.19	
(3.37)	(2.10)	(3.53)	(1.07)	(1.97)	(0.88)		
			Accounting NOB = 440				
2.90	-0.003	.0010	-.0001	-.0137	.06	.22	
(4.27)	(3.92)	(5.20)	(0.11)	(3.15)	(1.58)		

* The Sessional Dummy is 0 for Sessional, 1 for Regular Faculty. The Graduate Dummy is 0 for Graduate, 1 otherwise. The Area Variable is a polytomous variable which takes a value between 1 and 7, depending on the group. (A specific regression for International Business was not reported due to the small number of observations.) The Course Level variable is the actual course number, e.g., 232 or 312. The Year Variable is in the trimester form, 89.1, 89.2, 89.3, 90.1, and so on. Number in brackets below coefficient estimate is the absolute value of the t test for the null hypothesis that the coefficient equals zero.

Table 7

Multiple Regression Results for Teaching Responses*

Sample: NOB = 1600 (Includes All Courses and Sessionals)

Dependent Variable: Teaching Score (T)

Estimated Equation: $T = a_0 + a_1$ Class Size $+ a_2$ Course Level $+ a_3$ Area $+ a_4$ Year $+ a_5$ Graduate Dummy $+ a_6$ Sessional Dummy

a_0	a_1	a_2	a_3	a_4	a_5	a_6	R^2
			FACULTY				
1.44	-0.002	.005	.0006	.0016	-.05	.100	.08
(2.66)	(3.85)	(0.94)	(5.04)	(2.73)	(2.45)	(3.52)	
			Marketing NOB = 268				
-1.06	-0.001	.0002	.0044	.0122	.20	.07	
(0.78)	(0.33)	(0.73)	(2.86)	(0.23)	(2.98)		
			Human Resources Management NOB = 242				
0.54	-0.005	.0000	.0032	.0139	-.07	.11	
(0.42)	(4.28)	(0.09)	(2.21)	(0.31)	(0.93)		
			Business Policy NOB = 227				
-1.19	-0.063	.0003	.0051	-.0506	-.06	.09	
(0.70)	(2.81)	(0.86)	(2.65)	(0.87)	(0.66)		
			Management Science NOB = 263				
1.18	0.001	.0012	.0015	-.0478	.09	.19	
(0.95)	(1.51)	(5.08)	(1.10)	(1.22)	(1.39)		
			Finance NOB = 146				
5.55	-0.004	.0008	-.0031	-.1070	-.23	.19	
(3.67)	(2.67)	(2.42)	(1.81)	(1.78)	(1.25)		
			Accounting NOB = 440				
3.84	-0.004	.0009	-.0009	-.1265	.11	.13	
(3.69)	(3.07)	(3.11)	(0.77)	(1.90)	(1.99)		

* See Notes to Table 6.

NOTES

1. Remmers (1928, 1930) are references in the early literature. Useful surveys of available studies include: Feldman (1978), Aleamoni and Hexner (1980), Marsh (1984, 1987), Aleamoni (1987), Dowell and Jones (1982).

2. There is no reason to suppose that student learning, per se, is the best criteria for teaching effectiveness, e.g., Scriven (1989). For example, a high level of student learning of material which is not relevant to the requisite curriculum coverage is typically undesirable. In addition, for more advanced courses, it is not always apparent what the requisite curriculum should cover or whether problem solving techniques should be emphasized at the expense of other topics. A number of studies have examined the issue of identifying teacher effectiveness in a business school context, e.g., Etherington (1989), Kemp and Kumar (1990) and Tanner, et.al. (1992).

3. The connection between these results and the use of factor analysis should be apparent. For example, Feldman (1976) identifies twenty different categories relevant to teaching effectiveness. If the single factor model is correct, the bulk of the response items will load predominately on one factor. However, because different sampling instruments, e.g., SEEQ and IDEA, are designed to achieve measures of implicitly different teaching effectiveness criteria, it is possible that differing factor loadings may occur.

4. This evidence leads Gramlich and Greenlee (1993) to conclude: "If there is a message in all this, it seems to be that although something about instructors does matter in explaining student learning…that something is not well measured by the student evaluations of teaching." (p.4) While plausible, this conclusion may be confounded by studies such as Feldman (1976) which show little or no relationship between student ratings and grade point averages. In effect, student learning as measured by test or course grade performance may be dependent largely on student quality which is outside the control of the instructor, e.g., Watts and Bosshardt (1991).

5. Statistically, it is required that the omitted variables are not orthogonal to the included variables for global estimates to be biased.

6. Of the variables listed, the impact of grades has received the most attention. Considerable controversy has surrounded this topic, e.g., Aleamoni and Hexner (1980) and Aleamoni (1987) for a review. Recent research by Engdahl, et.al. (1993) has indicated that expected, as opposed to actual, grades are the appropriate variable to consider.

7. In identifying a significant, albeit lesser, impact for variables such as instructor knowledge, perceived fairness and student learning, Clayson and Haley (1990) also confirmed the multidimensional feature of student ratings.

8. The FBA offers second (200 level), third (300 level) and fourth year (400 level) undergraduate courses plus regular MBA (500 and 800 level) and executive MBA (600 level) graduate courses. Because not less than thirty credit hours are required for student admission to the business school, second year courses are available to non-business students.

9. Current practice in the SFU FBA is to keep all information contained in the student evaluations confidential, with the exception of information from the global teaching measure. At the end of each trimester, this data is released to individual faculty members in rank order, from highest to lowest score, without identifying individual instructors. Hence, each instructor is permitted to know only his relative rank on the global teaching response, without knowing what specific scores correspond to other instructors. This emphasis on the teaching score as the primary summative measure of teaching performance is embodied in the FBA tenure and promotion process. Letters from the faculty tenure and promotion committee (FTC) to faculty regarding FTC decisions about faculty progress typically make specific reference to the actual teaching scores received, using these scores

as the primary measure of faculty teaching performance.

10. Certain courses in accounting are an exception because its curriculum is tailored to meeting the requirements of external certification bodies such as the CGA and CA associations.

11. For example, empirical evidence from various studies indicates that there may be feedback between extraneous factors, such as expected grades, and observed responses. Stratton, et.al. (1994) find that the introduction of student teaching evaluations resulted in a significant increase in grades received.

12. One complication which arises in using analysis of variance in testing business school students is accounting for sample selection bias. This arises because, outside of core courses, the pool of students being sampled in each of the functional groups is not the same. If present, this effect could be reflected in the data in various ways, making its impact on specific null hypotheses difficult to predict.

13. A sessional instructor is defined according to the type of contract held by the instructor. More precisely, a sessional instructor does not hold either a tenure track or tenured position. In the FBA, a number of sessional instructors have had long term employment. The bulk of the courses taught by sessional instructors have been given by these long term sessionals. Hence, the discrepancies between sessional and regular faculty results cannot be attributed to learning effects of new instructors.

14. Attempts to provide empirical explanations for the behaviour of the difference were unsuccessful. More precisely, residuals were generated from a regression of teaching on course scores. These residuals were then regressed on variables used in this study, e.g., class size and course level. The constant was the most significant explanatory variable in these regressions.