A Model for Measuring Teaching Effectiveness using Student Evaluations

Ayodele Mobolurin
Associate Professor
School of Business
Howard University
Washington, DC 20059

Mohammad A. Quasem
Professor
School of Business
Howard University
Washington, DC 20059
(202) 806-1602

Charles M. Ermer
Associate Professor
School of Business
Howard University
Washington, DC 20059
(202) 806-1462

Abstract

This paper develops a model for including the results of student evaluations as part of an overall Faculty Evaluation Program. The approach is pragmatic in that the modeling effort avoids the impossible task of ranking all faculty members as a subset used to develop the model. This subset is tested using the Quade test to verify that ranking of the individual faculty members in the subset is justified. The model is regression based using the six summary scores obtained from Student Instructional Reports. The model is compared to published results of studies of the evaluation process.

Overview

The initial motivation for this study was the requirement for the School of Nursing of Howard University to develop a Faculty Evaluation Program during the 90-91 academic year. The plan called for each faculty member to be reviewed with respect to two categories; Teaching Effectiveness, and Scholarship and Service. The goal was to provide a numerical index, between 0 and 4, for each faculty member. This index would be a weighted combination of Teaching Effectiveness at 60% and Scholarship and Service at 40%. This summary index would allow the overall ranking of individual faculty members.

The Scholarship and Service score would be a number between 0 and 4 determined by a review between 0 and 4 determined by a review between the faculty person and the department chairperson based on the individual faculty members own predetermined objectives. The annual process of defining objectives for Scholarship and Service and assigning weights to these objectives determines the quantitative model under which each faculty member would be evaluated. The annual review would thus give each faculty member an informative review of their progress as well as a measure of their performance.

The Teaching effectiveness score would be a number between 0 and 4 that would be a weighted combination of the following teaching evaluations; Student, Peer, and Chairperson. The weights assigned to each evaluation are: Student, 45%, Peer, 25%, and Chairperson, 30%. The evaluation forms used for the Peers, and Chairpersons evaluations are based on a 0 and 4 scale and a summary score between 0 and 4 is provided. The student evaluation form used by the School of Nursing is the Student Instructional Report, SIR, of the Educational Testing Service.

While the SIR scores individual questions on a 0 to 4 scale it does not compute a composite score. The SIR gives a set of six summary factors; i.e. Factor, scores for the following topics:

1. Course Organization and Planning,
2. Faculty/Student Interaction
3. Communication
4. Course Difficulty and Work Load
5. Text Books and Reading
6. Tests and Exams

In addition, the computed score for each of the above categories is not in the 0 to 4 range. See Table 1(APPENDIX I) for the range of scores in each category and the number of questions that are combined for each category's, factor, score.

Past Work

Since the evaluation of faculty with respect to teaching effectiveness plays an important role in promotion, retention, and tenure decisions the topic of student evaluations has been extensively studied. Early work was focused on the instrument itself, the student surveys. Bending (3) addressed the issue of how to utilize the results of multiple single scales. Simple summing of the individual question results would obscure important information; on the other hand the combining of inter-correlated questions to develop fewer indices would lead to easier interpretation of the results. Bending found that three factors accounted for eighty percent of the variance in instructor ratings.

As the teaching evaluation instruments were applied to the measurement of instructor effectiveness several concerns surfaced. Of primary concern was whether or not the instruments were valid; that is, do student ratings correlate with teaching effectiveness. In addition, questions of bias arose; that is, were external and uncontrollable factors biasing the student ratings. Among the bias concerns were: affects of actual or expected grades, course difficulty, the personality/attitudes of the instructor and student, the perceived value of the course, and the impact of class size.
An excellent summary of these issues is contained in an article by McKeachie (9). McKeachie concluded that student ratings are a valid measure of teaching effectiveness. With respect to biasing factors the conclusions are that the value/difficulty of the course and the sex of the instructor have no effect on the ratings. Results with respect to faculty rank, student personality, and relative leniency in grading are inconclusive. Finally, class size, whether or not the course is required, and instructor personality do impact the ratings.
However, in a recent study by Agarwal, Gong, Mukherji and Turner (2), they concluded that there is no significance difference between average student evaluations scores for small and large classes or between undergraduate and graduate classes taught by the same faculty member.

Abrami, Perry, and Leventhal studied the impact of student and instructor personality characteristics on student ratings. They concluded that there is no significant correlation between student personality characteristics and student ratings. However, teacher characteristics are correlated with student ratings. They conclude that student ratings are the best used for classifying instructors rather than ranking instructors. Howard and Maxwell (8) studied the effects of grades and student satisfaction on student ratings. They concluded that grading leniency had a minor effect on ratings but that student motivation; that is, the students desire to take the course, had a significant impact on ratings and thus should be controlled. Hofman and Kremer (7) concluded that when students and instructors share common attitudes towards higher education the instructor is more likely to receive a higher rating. Overall and Marsh (10) studied the effect of time on ratings. In particular, they were concerned if ratings would change based upon experience following graduation. They had students re-evaluate the instructors one year after graduation. The author concluded that time does not affect the evaluations.

A comprehensive study of the correlation between student ratings and teaching effectiveness as measured by student achievement was done by Cohen (5). Cohen overcame one of the weaknesses of prior studies, small sample size, by using the results of the prior studies as the basic data for the analysis. As a measure of teaching effectiveness Cohen used the generally accepted notion that student learning is not

the most effective measure of student achievement. Thus Cohen used only studies that used student learning as the surrogate measure of teaching effectiveness. Based on these studies Cohen found a mean correlation of 0.47 between student ratings and student achievement, with a 95% confidence interval of .09 to .73. Cohen considered this to be as strong correlation; thus concluding that student ratings are a valid measure of teaching effectiveness. The student rating instrument results were subdivided into six dimensions of teaching. These were skill, rapport, structure, difficulty, interaction, and feedback. The correlations between student achievement and each these six factors was also studied. The mean correlations for skill, rapport, structure, difficulty, interaction, and feedback were found to be .50, .31, .47, -.02, .22, and .31, respectively. The 95% confidence intervals on skill and structure were in the positive range; that is, the intervals did not contain any negative correlations. Cohen (4, pg. 305) concluded that "Students do a pretty good job of distinguishing among teachers on the basis of how much they have learned. Thus, the present study how much they have learned. Thus, the present study lends support to the use of rating as one component in the evaluation of teaching effectiveness. Both administrators and faculty should feel secure that to some extent ratings reflect an instructor's impact on students."

One area that Cohen noted for future study was that few studies had been made on advanced courses. This raises that question of whether or not other influences, such as student interest, may have a more significant impact on student achievement in advanced courses. Briggs, Champion, and Gosenpud (4) studied an upper level required course in a school of business curriculum, "Production and Operation Concepts". They studied the correlation between the student responses and student achievement to the student response to two single questions relating to Best Professor and Best Course. They concluded that low correlations between student achievement and student response to Best Professor and Best Course indicate that the single question Best Professor should not be used in personnel decisions. Unfortunately, they did not report on the correlations between the key factors Skill and Structure and student achievement that were determined to be significant in the Cohen study.

The general conclusions that can be drawn from the prior research are:

1. Student evaluations are a valid measure of teaching effectiveness as defined by student achievement/learning.
2. Some external factors; in particular, class size, interest in taking the course, and instructor personality, do influence the student evaluations.
3. Two factors; that is subcategories of the student evaluations, skill and structure are strongly correlated with student achievement.
4. When teaching effectiveness is part of a personnel decision; e.g. appointment, promotion, or tenure, student evaluations should be only one part of the teaching assessment process. The student evaluation is a good measure of student achievement; however, evaluation of content, goals, and level of achievement should involve peer evaluations.

Problem

The problem faced by the school of Nursing was: how to use the information provided by the students via the SIR to evaluate Faculty Teaching Effectiveness? In particular, how to combine the six Factor scores into one 0 to 4 score?

Methodology

The model development utilized the results of the student evaluations from the 1989-90 academic year. The major difficulty is the lack of an absolute measure; that is, most common measure of student achievement, with which to correlate the results of the student evaluations. It was decided to select 6 faculty members from the population of 25. The criteria would be that these 6 could be assigned a score with respect to teaching effectiveness and that these scores would be consistent with the SIR data. The reason for using a subset of the faculty was to eliminate the need to rank all faculty members and then assign 0-4 ranking to each. This task would have required too fine a definition of individual scores between faculty members.

Before proceeding with the model development that would use scores provided by the Dean it was necessary to verify that the student evaluations, SIR data, could be used for individual comparisons of the six faculty. This test was performed using the nonparametric Quade test. Since the data passed the Quade test multiple regression would be used to develop the model. The Dean was then responsible to assign a teaching effectiveness rating, between 0 and 4, to each of the six. With the Dean's qualitative ratings, acting as the dependent variable and the 1989-90 academic year results as the independent variables a regression model for teaching effectiveness was developed. This faculty data was then tested by performing a ranking of all faculty members for the 1989-90 academic year data. The full ranking was then reviewed to see if the results of the ranking were in general consistent. It should not be expected that the ranking would be precise but that it is generally accurate and fair.

Analysis

The Dean selected six faculty members that could be clearly distinguished with respect to Teaching Effectiveness. The results from the 89-90 academic year were collected for each and the Dean assigned a Teaching Effectiveness score. The 89-90 SIR results for the six are shown in Table 2. The Dean's ratings were: 3.75, 3.75, 1.75, 2.50, 3.25, and 3.00 for faculty A, B, C, D, E, and F; respectively.

The Quade Test

The Quade test, Conover (6), is a nonparametric test that uses rankings to detect differences in multiple subjects based on several related samples of an experiment on each subject. A nonparametric test is appropriate for the initial analysis of the data in Table 2 since this test does not require that the data be normally distributed. The objective of the test in this case was to determine if the results of the Factor scores could be used to detect differences in the individual faculty members. The Factor scores are thus the blocks and the faculty are the treatments in the test. The process begins by ranking each faculty based on their score on each Factor. The results are shown in Table 3. The rankings are based on the average score for each faculty on each Factor; the X6 score for faculty member A was the average of the X1-X5 scores for faculty A. Next the Factors (Blocks) are ranked based on the range of each factors scores, and a statistic Sij,

that represents the relative overall ranking of each entry is computed. The results are shown in Table 4. The first step in the Quade test is to test the Null Hypothesis: The rankings of the faculty within the Factors (Blocks) are equally likely. The test statistic is given by:

$T1=(b-1)B1/(A1-B1)$

1: Where b is the number of blocks and A1 and B1 are given by

2: $A1=\sum \sum Sij2$

3: B1: $(1/b) \sum Sj2$ $\quad B1 = (1/b) \sum Sj2$

Performing the calculations yields, T1=13.31. The threshold for the .95 quartile with k1=k-1=6-1 and k2= (b-1) (k-1)= (6-1) (6-1)=25 is 4.53, where k is the number of faculty (treatments), see F Distribution Table A26, Conover (6, pg. 483). Thus the null hypothesis can be rejected and multiple comparisons can be made between faculty members using the Factor score rankings.

Thus the next step was to make pair-wise comparisons of the six faculty. Two faculty are considered different if the inequality

4. $Si-Sj > t1-a/2 \, [2b(A1-B1))/(b-1) \, (k-1)]^{1/2}$

Where t1-a/2 is obtained from the t Distribution with 25 degrees of freedom for this case, is satisfied. The right hand side of the inequality is 29.82. The result of the pair-wise comparisons is shown in Table 5. Thus it has been shown that; in general, the Dean's choices of the Faculty can be ranked, and that the data from the SIR reports supports this ranking. The only issue in question is that the rankings of the adjacent pairs of faculty can not be made; at least not at the a=.05 level. With respect to this, two points must be made; first, the Quade test uses rankings, thus information on the relative distance between Faculty as provided by the SIR score is lost. Second, in some cases the Dean's qualitative scoring reflects this inability to recognize differences; for example, both A and B are scored at 3.75. On the other hand, faculty member D and C can be ranked if a is increased to .10, this is consistent with the Dean's ratings of 2.5 and 1.75 for D and C, respectively. Thus the nonparametric analysis has demonstrated that the faculty members chosen can be ranked on the basis of SIR scores.

The Regression Model

A correlation analysis was performed to test for collinearity, the results are shown in Table 6. The results for the complete data, X1-X6, are based on 13 data samples since in some cases data is missing due to insufficient responses to the questionnaire. As could be expected collinearity is significant. In order to see the effect of using the complete data, i.e. all 20 data points, X5 and X6 were dropped. Table 7 shows the correlation of X1-X4. As expected, the variables exhibit high collinearity and the results are similar to Table 6.

Due to the high degree of collinearity it was decided to use step-wise regression to develop the model. Using the Dean's ratings as the dependent variable and the X1-X6 factor scores as the independent variable a step-wise regression was performed. For the step-wise regression the parameters were: significance level to enter was .05, significance level to leave was .10, and cut off value for tolerance was .001. This analysis produced a simple regression model that was based on X1 alone. The model had an R-squared value

of .911, with an F value of 113.95. Since this model used only 13 data points, due to missing data, another step-wise regression was run. This second model used the Dean's ratings as the dependent variable and the X1-X4 factors as the independent variables. Since there are no missing data this model was based on all 20 data points. Again the model was a simple regression model,

Y= -0.03125+0.31685 X1

For this model the R-Squared value is 0.847, the F value is 100.03, and the Standard Error of the Estimate is .28.

The next step was to compare the model's ratings to the Dean's ratings. Table 8 shows that there is an inconsistency between the Dean's ratings and the student ratings. The student ratings reverse the ranking of faculty E and F. This reversal is caused by faculty F having a higher average score on Factor X1 than faculty E. Since the purpose of the effort is to develop a student rating model it was decided to let the student preference take priority over the Dean. Thus the Dean's ratings for E and F were swapped. Based on the new dependent variables a new regression model was developed. The new, and recommended model is

Y=-.12873+.32680 X1

For this model R-squared is .902, the F value is 164.78, and the Standard Error of the Estimate is .225. This model was again based on step-wise regression using Factors X1-X4 as the dependent variable. With a threshold value of 26.87 on the F value for an α of .01 the model is a very good fit.

The 89-90 results of the student evaluations of the faculty were collected. This consisted of 44 data points, again some data points were incomplete with respect to X5 and X6. The model given by equation (6) was applied to the full faculty, where an individual faculty member taught more than one course the average Student Teaching Effectiveness Rating for the courses was computed. The distribution of the scores is shown in Table 9. Based on a review of these scores it was felt that the model adequately represented the ranking of the faculty from the student's perspective. Excluding the single lowest score the range of the scores was 1.12 which is 4 Standard Errors of the Estimate. Four distinct groupings emerged, that is gaps between adjacent scores were approximately one Standard Error of the Estimate. These groupings were: 3.59-3.75 with 4 faculty, 2.85-3.34 with 15 faculty, 2.63-2.65 with 2 faculty, and below 2.0 with one faculty member. Since these results are what would be anticipated by any statistical model it is felt that the model is representative of the student evaluations.

Conclusions

A model was developed that allows the results of the student evaluations to be included as one element if an overall faculty evaluation program. In order to be able to clearly distinguish between individual faculty in developing the model a subset of faculty members was selected for the analysis. This avoided the impossible task of attempting to rank and assign scores to every faculty member. The resulting data was tested to verify that rankings could be made by using the Quade test. A regression model was developed that assigns a 0-4 score for each faculty member based on the results of the student evaluations. The model was found to be consistent with existing research; that is, the model is based on the variable organization and planning. Structure, which is the

equivalent of organization and planning, was found by Cohen to be highly correlated with student achievement. Finally, the overall evaluation program uses inputs from both peers and chairpersons thus the program meets the recommendation that student evaluations not be the sole source of data for a teaching evaluation program.

References

1. Abrami, P.C., Perry, R.P. and Leventhal, L. "The Relationship between Students Personality, Teacher Ratings, and Student Achievement", Journal of Educational Psychology, 1982, Vol. 74, No.1, pgs. 111-125.
2. Agarwal, K., Gong, W., Mukherji, S., and Turner, C., "An Analysis of Student Evaluations of Faculty Teaching", Presentation, Howard University, May, 2012.

3. Bending, A.W., "A Factor Analysis of Student Ratings of Psychological Instructors on the Purdue Scale", The Journal of Educational Psychology, 1954, Vol. 45, No. 7, pgs. 385-393.
4. Briggs, J.R., Champion, W.M., and Gosenpud, J.J., "The Student Evaluation of Instructors in Required Operations Management Courses", Operations Management Review, 1990, Vol. 8, No. 2, pgs. 14-30.
5. Cohen, P.A., "Student Ratings of Instructors and Student Achievement: A Meta-Analysis of Multisection Validity Studies", Review of Educational Research, 1981, Vol. 51, No. 3, pgs. 281-309.
6. Conover, W.J., Practical Nonparametric Statistics, 2nd Edition, John Wiley and Sons, New York, (1980).
7. Hofman, J.E. and Kremer L., "Attitude Toward Higher Education and Course Evaluation", Journal of Educational Psychology, 1980, Vol. 72, No. 5, pgs. 610-617.
8. Howard. G.S. and Maxwell, S.E., "Do Grades Contaminate Student Evaluation of Instruction?", Research in Higher Education, 1982, Vol. 16, No.2, pgs. 175-188.
9. McKeachie, W.J., "Student Ratings of Faculty: A Reprise", Academe, October 1979, pgs. 384-397.
10. Overall, J.U. and Marsh, H.W., "Students' Evaluation of Instruction: A Longitudinal Study of their Stability", Journal of Educational Psychology, 1980, Vol. 72, No. 3, pgs. 321-325.

APPENDIX I

Table 1: SIR Categories

Factor Categories	Number of Questions	Factor Score Range
X1-Organization & Planning	7	3.73-12.22
X2-Faculty/Student Interaction	8	4.44-12.39
X3- Communication	6	4.59-13.03
X4- Course Difficulty & Work Load	3	4.17-12.48
X5-Textbooks & Readings	2	2.93-13.42
X6-Test & Exams	2	4.26-13.00

Table 2: Baseline Data

Faculty	X1	X2	X3	X4	X5	X6	Class Size
A	11.70	12.38	9.02	11.36	10.49	N/A	8
	11.94	12.38	11.45	10.93	N/A	N/A	5
	11.37	12.20	10.44	10.74	N/A	N/A	8
B	12.04	12.38	10.81	11.51	11.96	12.34	19
	11.59	12.38	8.04	12.24	11.79	12.01	11
	12.00	12.12	11.04	10.11	12.79	N/A	5
	11.85	12.38	11.14	10.71	12.48	N/A	5
C	5.88	6.10	7.88	8.03	10.67	7.74	22
	6.09	6.64	8.46	8.70	9.84	7.57	26
	6.65	7.83	8.38	9.43	N/A	7.96	N/A
D	8.86	9.18	8.78	8.53	11.17	9.19	22
	8.09	8.09	9.24	10.96	9.43	8.03	16
E	9.61	9.46	9.64	10.55	9.81	9.34	11
	9.43	9.22	8.81	8.90	11.48	9.58	15
	7.55	7.97	7.98	9.23	N/A	7.32	N/A
	10.86	10.78	9.43	10.27	10.91	11.16	37
F	9.56	9.58	9.23	11.12	11.24	10.85	21
	10.17	10.16	8.57	9.58	10.93	9.37	14
	10.19	10.72	8.68	9.58	9.42	10.98	9
	10.64	10.56	8.72	10.63	0.48	10.84	N/A

Table 3: Faculty rank within block

Factors (Blocks)	A	B	C	D	E	F
X1	5	6	1	2	3	4
X2	6	5	1	2	3	4
X3	6	5	1	4	3	2
X4	5	6	1	3	2	4
X5	3	6	1	2	5	4
X6	5	6	1	2	3	4

Table 4: Ranking Statistics

Factors (Blocks)	Factor Rank Q1	A	B	C	D	E	F
X1	6	9	15	-15	-9	-3	3
X2	5	12.5	7.5	-12.5	-7.5	-3	3
X3	2	5	3	-5	1	-1	-3
X4	3	4.5	7.5	-7.5	-1.5	-4.5	1.5
X5	1	-0.5	2.5	-2.5	-1.5	1.5	0.5
X6	4	6	10	-10	-6	-2	2
		36.5	45.5	-52.5	-24.5	12	7

K= the number of treatments= 6

Table 5: Pair-wise comparisons

Ranked Faculty	A	F	E	D	C
B	NS	S	S	S	S
A		NS	S	S	S
F			NS	S	S
E				NS	S
D					NS
NS=not significant S=significant					

Table 6: Correlation matrix for SIR data

Factor	X1	X2	X3	X4	X5	X6
X1	1.000	0.986	0.478	0.713	0.478	0.924
X2		1.000	0.423	0. 720	0. 485	0. 943
X3			1.000	0.445	0.154	0.411
X4				1.000	0.246	0.695
X5					1.000	0.543
X6						1.000
Note: Due to missing data this correlation is based on 13 out of 20 samples.						

Table 7: Correlation matrix for abridged SIR data

Factor	X1	X2	X3	X4
X1	1.000	0.983	0.677	0.723
X2		1.000	0.663	0.737
X3			1.000	0.444
X4				1.000
Note: This correlation uses all 20 samples.				

APPENDIX II

Table 8: Different Course Types (from Agarwal, Gong, Mukherji and Turner (2))

	Mean Diff.	T-stat
	No. of Students	
Large - Small	25.4**	21.7
Ugrad - Grad	7.2**	3.1
BCore – NBCore	10.5**	5.3
	Eval. Score	
Large - Small	0.04	0.43
Ugrad - Grad	0.17	1.41
BCore – NBCore	0.04	0.41
**Significant at 1% level.		
*Significant at 5% level.		