STUDENTS' EVALUATIONS OF INSTRUCTORS: BEFORE AND AFTER THE EXAMINATION, NAMES IDENTIFIED *VERSUS* ANONYMOUS

LES ÉVALUATIONS DES ENSEIGNANTS PRÉPARÉES PAR LES ÉTUDIANTS: AVANT ET APRÉS L'EXAMEN, L'ÉTUDIANT S'IDENTIFIE/L'ÉTUDIANT RESTE ANONYME

Is there any difference between students' before-the-examination evaluations of instructors and their after-the-examination evaluations, and between students' name-identified evaluations and their anonymous evaluations?

There has been a great deal of controversy, both theoretical and empirical in nature, relative to the issue of students' evaluations of insturctors. On the negative side of the theoretical dimension, Cronbach (1966), Jackson (1968), and Broudy (1969) hold the general view that there is insufficient knowledge about the learning process to enable us to assess instruction effectively. On the positive side which is heavily predicated upon empirical evidence, Hamachek (1969) contends that competent, effective or successful teaching can be identified if systematic attempts were made to identify these different specific skills. Mcdonald (1978), however, examines both sides of the issue on teaching effectiveness. He states that in order to obtain valid data on specific skills, knowledge, and attitudes, all instructors are required to be trained. The type of training that he proposes is divided into three levels, namely, modular training on a specific component skill, intermediate training for integrating these component skills into a coherent set of activities throughout an instruction period, and practice teaching over an extended period of time, such as several weeks. He also cautions us that it is difficult or impossible to identify what teachers do that makes a difference in children's learning.

It is generally accepted that a considerable amount of data pertaining to the reliability of students' ratings of instructors has been collected. More recently Oles (1975) found there was a significant correlation between students' pre-instruction evaluations of instructors and their post-instruction evaluations, and he further stated that students generally gave a lower rating to their course and the instructor at the end of the semester than they were at the beginning, but he did not speculate on the reasons since it was not the purpose of his study to do so. Therefore one of the concerns of my study was to find out the difference, if any, between students' evaluations of the instructor before and after the students received marks in the one-semester courses.

Validity data on students' evaluations of instructors are available. Costin, Greenough, and Menges (1971) reviewed a number of studies in which the final grade was used as the criterion; they summarized that approximately half of the studies showed positive but low correlations between students' grades and their evaluations of instructors, while the other half showed no or negative relations. Follman (1975), in a systematic review of studies dealing with the extent to which student ratings of instructors' teaching effectiveness are

* Published by Ontario Confederation of University Faculty Associations, 40 Sussex Avenue, Toronto, Ontario, Canada M5S 1J7.

influenced by characteristics of the raters, concluded that ratings are substantially influenced by rater's personality characteristics.

The purpose of my study was to investigate the following hypotheses:

1. Students' name-identified course evaluation scores of an instructor's teaching effectiveness are significantly higher than their anonymous evaluation scores.
2. Students' anonymous course evaluation scores of an instructor's teaching effectiveness before course marks given are significantly higher than those after course marks given.
3. Male and female students do not differ in their course evaluations of instructors when they identify themselves and when they remain anonymous.
4. M.A. students and B.Ed. students do not differ in their evaluations of instructors when they identify themselves and when they remain anonymous.

Method

Forty-two Bachelor of Education degree students who took two half-unit courses and twenty M.A. students who also took two half-unit courses were the subjects of this study. It should be mentioned that there were reductions in numbers in several testing sessions since they were absent during particular class periods.

Prior to the presentation of the data collection process, a brief description of instrumentation appears to be in order.

The administration and the faculty of Mount Saint Vincent University (perhaps a majority of them) have professed to the public that the pursuit of excellence in teaching is our goal. Obviously such an enunciation is well in accord with the current trend.

Subsequently, the Senate of the University appointed an *ad hoc* committee during 1973-74 academic year in order to develop a questionnaire for appraising instructors' teaching effectiveness. There were two open meetings held to discuss, perhaps as a matter of formality, the proposed instrument, in which objections were voiced by many faculty members. But no votes were taken at those open meetings. The manner in which those open meetings were held was perhaps intended to be manipulated in such a way that the approval to implement or not to implement the formal evaluation of instructors' teaching effectiveness did not rest with the faculty. After two open meetings, the questionnaire consisting of 30-items was taken back to the Senate where it was approved, and the decision to implement formal evaluations of teaching of all instructors was made by the Senate.

Therefore, all faculty members of the University were required to have formal course evaluations in 1974-75 academic year. Realizing the opportunity to undertake a study of the validity of students' evaluations of instructors, I administered this identical instrument twice to my M.A. and B.Ed. classes respectively, one in early December of 1974 before the final examination of the one-semester course, when the students were asked to remain anonymous by the investigator, and once in early March of 1975 when the students were asked to identify themselves. But for the latter administration of the instrument, students were assured by me that I would not open the sealed envelopes for analysis of a research project until course marks were submitted in May of 1975, and that the sealed envelopes

would be available for inspection at any time in my office in order to make sure that there was no breach of promise.

In mid-April of 1975 the Committee on Course Evaluation formed by the University Senate administered the questionnaire (here-after known as the official, anonymous evaluation) to my classes and collected and analyzed the same. Since these data were related to the research project in question, I requested in advance the Chairman of the Committee to return, after their analyses, the anonymous answer sheets to me for research purposes. My request to these two groups of students three times, twice anonymously and once in which they were asked to identify themselves. To put it in another way, the instrument was administered once *before* they received course marks for the one-semester course, and twice *after* they received the course marks.

Results, Interpretations, and Implications

Since the B.Ed. and the M.A. groups were separate, data analyses were performed within each group respectively.

Hypothesis No. 1 was concerned with the difference between one set of students' evaluations in which they were asked to identify themselves and two sets of anonymous evaluations of the instructor.

Table 1 presents the number of *S*s, Means, and standard deviations of both the B.Ed. and the M.A. groups' two anonymous evaluations and one name-identified evaluation of the instructor (insert Table 1 here) for the scores on the entire instrument consisting of 30 items as well as for the scores on the last two items of the instrument, namely

Overall, this Professor was very poor (1) to excellent (9).

Overall, the course was very poor (1) to excellent (9).

In the first instance, a statistical test (Downie & Heath, 1974) was applied to data collected for anonymous evaluation in December of 1974 (before course marks given) and for evaluations in which names were identified in March of 1975, for the B.Ed. group, the result was not statistically significant; for the M.A. group, the result was not statistically significant. Therefore, for scores on the entire instrument of both groups, this research hypothesis was rejected. These findings indicated that there was no significant difference between students' anonymous evaluations, (before course marks given) and their evaluations (after course marks given) in which students were asked to identify themselves. Meanwhile, analyses were also made to their mean scores on the last two items of the instrument, the results obtained for B.Ed. and M.A. groups respectively, were not statistically significant. This finding reinforced the preceding interpretation. It might be appropriate to speculate here that in spite of my assurances, students in both groups perhaps felt apprehensive about giving me a significantly lower rating even after they recieved the course marks because they had to identify themselves.

Then analyses were made to data collected on name-identified evaluations and on the official, anonymous evaluations. The results obtained for the B.Ed. and M.A. groups respectively, were statistically significant. In this case the research hypotheses was therefore accepted. This finding meant that students' evaluations in which they were asked to identify themselves were significantly larger than the official, anonymous evaluations

TABLE 1

Means and Standard Deviations of

Students' Evaluations of The Instructor

Categories	B.Ed. Group			M.A. Group		
	N	$\bar{X}$	s	N	$\bar{X}$	s
Scores of The Entire Instrument						
1) Anonymous Evaluation before Marks given in December 1974	42	163.88	27.40	20	181.05	13.47
2) Names Identified in March 1975 after Marks given	37	156.08	21.89	18	180.17	13.61
3) Official, anonymous Evaluations in April 1975	41	152.30	19.50	19	175.00	19.00
Scores on The Last Two Items of The Instrument						
1)	42	13.02	3.43	20	15.30	2.00
2)	37	12.68	3.43	18	15.94	1.66
3)	41	11.83	3.84	19	15.37	1.64

which were handled by the University committee. This finding also indicated that regardless of the academic level, i.e., whether it be at the B.Ed. or the M.A. level, students felt apprehensive about giving me a low rating when they had to identify themselves. In a sense this finding confirmed the speculation offered in the preceding paragraph. Analyses were also made on the mean scores of the last two items; the result obtained for the B.Ed. group was statistically significant, and thus provided additional support for the conclusion here.

Hypothesis No. 2 was related to the difference between students' anonymous evaluations of the instructor in December, 1974, that is *before* they received course marks for the spring semester course, and the official students' anonymous evaluations of the instructor in April of 1975, that is *after* they received marks for the fall semester course.

The result for the B.Ed. group was highly significant, but for the M.A. group, there was difference, yet not statistically significant. This finding indicated that for the B.Ed. group, their anonymous evaluations of the instructor were significantly smaller after they received course marks than evaluations that were made before course marks were received. However, this lower rating by students of the instructor could have tremendous practical significance for the University Administration in making decisions relating to salary increase, promotion, and the granting of tenure, which will be discussed later in this paper.

Hypothesis No. 3 was concerned with sex difference relative to male and female students' evaluations of the instructor. This item of information on sex was available of two sets of data, namely anonymous evaluations in December of 1974 and name-identified evaluations in March of 1975. The results obtained were not statistically significant. The null hypothesis was therefore accepted.

Hypothesis No. 4 referred to the difference, if any, between the B.Ed. group (prospective teachers) and the M.A. group (experienced teachers). Comparisons were made on three sets of data, namely anonymous evaluations in December of 1974, name-identified evaluations in March of 1975, and the official, anonymous evaluations in April of 1975. The results for anonymous evaluations in December of 1974 for the entire instrument and that for the last two items on the instrument, were both statistically significant; the result for the students' evaluations in which they were asked to identify themselves in March 1975 for the entire instrument and that for the last two items, were statistically significant. The results for the official, anonymous evaluations in April of 1975 for the entire instrument as well as for the last two items of the instrument, were again statistically significant. These findings meant that the M.A. group (experienced teachers) gave me a significantly higher rating in all cases than the B.Ed. group (prospective teachers). Why was this so? It appears reasonable to speculate that usually the experienced teachers would be more rigorous in judging an instructor's teaching performance than neophytes. However, in this particular case the converse was true. How do we explain such a phenomenon?

The University Administration had announced in the fall of 1975 that the marks for 1974-75 academic year were too high. Consequently the administration decided to tabulate marks for all courses of the University for distribution. I therefore obtained a copy of the course marks; then I was able to compare my marks with those of other instructors teaching the same group of B.Ed. students. In one case an instructor gave an average mark of 81 per cent to the class; whereas I gave an average mark of 65 per cent to the same class. A statistical test was performed, and the result was statistically significant beyond .001 level. Of course, I had no access to students' anonymous evaluations of that particular instructor. It would certainly be interesting to find out if students gave this particular instructor a significantly higher rating than what they gave me because we both taught the identical group of students. However, since I was given the departmental profile, I knew that my students' evaluations were below the departmental average. I believe that most likely this is the manner in which the University Administration would use these students' evaluations, that is, by comparing all faculty members (norm-referenced evaluation) instead of examining each faculty member's merits (criterion-referenced evaluation which is a more humane and more rational model as accepted in educational measurement today). For the M.A. group, the situation was different, because it was not possible to get an identical group of students taught by two different instructors.

On the basis of the statistically significant findings of the study, namely (i) students' course evaluations in which they were asked to identify themselves were significantly higher than the official, anonymous evaluations handled by the University committee, (ii) the B.Ed. group's anonymous evaluations were significantly lower after they received the course marks than those before they received the course marks, and (iii) the M.A. group's ratings were significantly higher than those of the B.Ed. group, we can draw several inferences:

Firstly, when I examine the Statement of Philosophy and Objectives of this institution, I find that it says that "Mount Saint Vincent University emphasizes excellence in teaching. To serve the cause of good university teaching and as a preparation for it, the faculty engage in research and scholarly activity. In addition to this basic research for teaching, the search for new knowledge and the adaptation of the old to the new are distinguishing features of the University." We are a small liberal arts and science college with selected professional departments, plus two small graduate programmes at the masters level. These are the "givens", and then I ask: Are our students competent to judge teaching effectiveness, as proclaimed in its objectives? Are students' evaluations valid indications of teaching effectiveness? What would happen to academic excellence that we at the universities have often proclaimed to pursue?

Secondly, students' evaluations of instructors are, I think, dangerous weapons, dependent upon the kind of person who is empowered to use them, for good as well as for evil. If they happen to be in the hands of an irrational university administrator, in the absence of other indications on teaching effectiveness, the essence of academic freedom, for which the universities have been fighting for centuries, is greatly in danger. Broader than academic freedom in scope is social justice which is the foundation of the western democracy; it would be, to say the least, at stake. This is because we all are human beings, and so we are frail.

Thirdly, even though the instrument used in this study consisted of thirty items, it is very easy, perhaps highly probable for a university administrator to go directly to the last two items, namely, the overall performance of the instructor, and the overall assessment of the course. Unless we can demonstrate that the whole is equal to the sum of parts, it is irrational and meaningless to use, not even to begin with, this kind of overall appraisal as a valid indication of teaching effectiveness.

Fourthly, as human beings, both the instructors and students are frail. When students perceive the importance of their evaluations of instructors, they or at least some of them might seize this opportunity, as it usually happens, for reprisal for whatever reasons known to them. On the other hand, when instructors realize the importance of these students' evaluations regarded by the University Administration, they or at least some of them might try to manipulate the students by means of high marks or some other methods available in their repertoire of strategies. Consequently, it is possible that university teaching, to a certain degree, then becomes an activity of mutual manipulation. If that happens, the real meaning of the university's existence is questioned.

It is appropriate to end this paper with a caution that the degree of generalizability of the study is limited on two accounts: the number of subjects used was small, and the investigator is a member of the visible minority group. Nonetheless it is hoped that further studies will be undertaken with larger samples and in other settings.

REFERENCES

Broudy, H.S., "Can We define Good Teaching?" *Teachers College Record,* 1969, 70, 583-92.

Costin, F., Greenough, W.T., and Menges, R.J., "Student rating of College Teaching: Reliability, Validity, and Usefulness," *Review of Educational Research,* 1971, 41, 511-535.

Cronbach, L.J., "How Can Instruction be Adapted to Individual Differences?" In R.M. Gagné (ed.) *Learning and Individual Differences.* Columbus, Ohio: Merrill, 1967.

Downie, N.W., and Heath, R.W., *Basic Statistical Methods.* (4th edition) New York: Harper and Row, 1974.

Follman, J., "Student Ratings of Faculty Teaching Effectiveness: Rater or Ratee Characteristics," *Research in Higher Education,* 1975, 3, 155-167.

Hamachek, Don., "Characteristics of Good Teachers and Implications for Teacher Education," *Phi Delta Kappan,* 1969, 50, 341-45.

Jackson, P.W., *Life in Classrooms.* New York: Holt, 1968.

McDonald, F.J., "Evaluating Preservice Teachers' Competence", *Journal of Teacher Education,* 1978, 24, 2, 9-13.

Oles, H., "Stability of Student Evaluations of Instructors and their Courses with Implications for Validity," *Educational and Psychological Measurement,* 1975, 35, 437-445.

George S.C. Cheong
Education Department,
Mount Saint Vincent University