STUDENT EVALUATIONS: HOW VALID AND HOW RELIABLE?

Allen J. Wedell, Colorado State Universtiy, Fort Collins, CO 80523, (303) 491-5155
Linda R. Stanley, Colorado State University, Fort Collins, CO 80523, (303) 491-7297

ABSTRACT

A pilot study of the effects of anonymity on student evaluations combined with a review of literature of student evaluations reveals major concerns regarding their validity and reliability. The authors suggest that if teaching effectiveness is to be measured, measurement must be obtained from relevant populations that are in a position to measure what has been learned and its relationship to the work world.

INTRODUCTION

Student evaluations have become widely used during the past two decades since students first exerted pressure on administrators to have greater input into university decisions. Now, student ratings are usually either required or highly encouraged by most institutions as a major criteria for faculty evaluation for salary, tenure, and promotion.

While many writers have studied student evaluations, evidence of their validity and reliability remain inconclusive. Investigators have studied such areas as instrument and sampling design, student characteristics and background, students' attitudes toward evaluation, instructor behavioral traits, and student grade point average. At best, the research has shown inconsistent results. Several studies report a negative or low correlation between student learning and the students' evaluation of the instructor. Rodin and Rodin [1972] found a negative correlation of .75 between these factors and Shamanske [1988] reported a low correlation in a longitudinal study of the same students enrolled in a subsequent class with the same instructor. Yet, the validity of student evaluations is claimed by Sullivan and Skanes [1974], Baird [1987], Marlin and Niss [1980], and Marsh [1977]. These studies report positive correlations between student learning and the student evaluations.

Other studies suggest that student evaluations do not measure what is sought to be measured, i.e., teaching effectiveness for student learning. Findings from these studies suggest that the evaluation process is influenced by outside variables [Dowell and Neal 1982] and that the evaluation instruments are ambiguous [Wheeler and Geurts 1986]. Therefore, an evaluation process does not reliably distinguish one departmental faculty from another. Sherman and Blackburn [1975] suggest the problem is enhanced by a lack of agreement on the criteria used for measuring teacher effectiveness.

A number of studies have revealed that students enter a class with opinions already formed or that they develop opinions of a class and the instructor very early in the term and that the subsequent learning has little influence upon that opinion [Sauber and Ludlow 1988] [Ortinau and Bush 1987] [Dowell and Neal 1982]. Orsini [1988] determined the presence of a halo effect suggesting that the questions on an evaluation form may not be taken at face value. He also states that faculty may be at a loss to improve a lowly rated teaching characteristic such as "fairness" or "ability to communicate clearly" as the students' perception is heavily influenced by other unmeasured factors.

One recent study involved the ranking of attributes students cited as most important for learning. Students ranked at the top such factors as "ability to motivate me," "an awareness of what students want to learn," "shows personal interest in me," "is a fun person," and "available outside of class." Ironically, a professor's "in-depth knowledge" ranked 10th and "evidence of scholarly accomplishment" ranked 15th, while "competitive grades" was at the very bottom [Eckrich 1990]. In another study of student evaluations, top concerns all centered upon exams, i.e., "tests related to course material," "fair test," and "indicates importance for exams," which suggest more concern about grading than about learning [Eagle & Catalanello 1987].

Studies showing a high positive correlation between "personality" and "teaching effectiveness" suggest that a professor wishing to improve his/her perceived effectiveness may do well to concentrate on personal attributes rather than on the course itself as students rate "warm" personalities as more effective instructors than "cold" personalities [Widmeyer and Loy 1988].

Another study revealed that the rigor of a course had a large and negative impact on students' perception of fairness [Clayson & Haley 1990]. In turn, students did not relate this fairness to the instructor's knowledge nor interest but rather to the instructor's personality. Thus, according to the study personality was an overwhelming determinate of a student's total evaluation. The presence of personal factors is further extended to gender as Baslow and Silberg [1987] found that students rate women differently than men.

The amount of literature on this subject is almost endless and the majority of it suggests low validity and reliability. It does, however, portray the presence of "many unknown factors." And for some time, researchers have believed these to be highly related to the personality of the instructor.

PILOT STUDY

While the numerous studies regarding student evaluations have investigated many factors, anonymity has not been addressed. To administer without accountability a process that has grown to have such impact on one's career seems to be a pursuit more for ease and convenience than one for valid and reliable measurement of teaching effectiveness. Limiting judgement of teaching effectiveness to students whose immediate interests are self-oriented and void of accountability does not provide an accurate nor credible evaluation of the teaching process [Eckrich 1990]. This is an important factor since policies concerning whether unsigned evaluations will count in assessing an instructor's teaching effectiveness differ from college to college. At some universities, all evaluations are used; at others, only signed evaluations are considered; and at some, signed evaluations are given more weight than unsigned evaluations.

Thus, a pilot project was undertaken during Spring Semester 1991 by one of the authors to examine how manipulating this factor in the process of administering student evaluations would affect the evaluation results. Each of three classes taught by the instructor (two sections of professional selling and one of sales management) was administered class evaluations twice during one class period. The first time the regular university procedure was followed. The second time the students completed the evaluation form, they were instructed to sign their name. They were assured that the results of the evaluations would not be seen by the instructor until after grades were submitted to the registrar's office. This was to minimize any positive effects on the signed evaluations due simply to students' fears that grades would be affected by the responses to the evaluations.

There are several hypotheses regarding the differences between the signed evaluations and the unsigned evaluations. Systematic differences may occur in the results so that the signed evaluations show either overall higher or lower scores than the unsigned evaluations. According to the study done by Eckrich [1990], 17% of students reported consciously using evaluations as a means of revenge on a professor. If students are held accountable to their evaluations, this tendency to seek revenge through evaluations is likely to be less, and thus, signed evaluations would, on average, be higher than the unsigned evaluations. This would be especially true for those evaluation items that pertain mainly to the professor, not specifically to the course.

In addition, students may consider their responses more carefully when they are held accountable to their evaluations, thus considering a larger set of evidence from the course in making their evaluations. This hypothesis suggests that experiences in the class that are more likely to affect an unsigned evaluation are those that are salient to the student. At the end of a semester, many specific class experiences will be forgotten. On the other hand, grades are soon to come. Salient experiences at the end are then likely to be those associated with grades in the class. If a professor is a "tough" grader, these salient experiences are more likely to be negative. On the other hand, if the professor is an "easy" grader, these salient experiences are more likely to be positive. A positive effect from signing the evaluations would be more likely in this study since the professor has a reputation for "toughness" and typically has the lowest grade average for his classes in the department. In addition, signing the evaluation may negate some of the tendancy found by Dowell and Neal [1982] whereby students only consider early opinions toward the instructor in evaluating that instructor. Signing the evaluation may motivate students into considering their overall learning experience in the class causing the signed evaluations to differ from the unsigned evaluations.

Second, random errors may occur if students are not involved while filling out the evaluations.

Eckrich [1990] found that 62% of students believe that evaluations are not used by the professor or the department in evaluating the professor. Therefore, there is little effort expended on them. Although students were given the evaluations during the same class period, if little effort was expended on them and if there are enough items on the evaluation so that the prior evaluation of each item will be difficult to remember, then some random differences between the two evaluations may occur.

Mean scores for the experimental and control groups (based upon a 16-item evaluation) of each class are presented in Table 1. The highest possible score for a single item is a 5 which indicates a strong agreement with a positive statement. In all but 4 cases (8%), the mean score for the signed evaluations is higher than that for the unsigned evaluations. T-tests to test for significant differences between the control and experimental groups is not possible since the samples are not independent and a match between each individual's signed and unsigned evaluations is impossible. Regression analysis is used, instead, to test for systematic and random differences between the signed and unsigned evaluations.

TABLE 1
Mean Scores From Student Evaluations

	Professional Selling, S1		Professional Selling, S2		Sales Management	
Item	Unsigned	Signed	Unsigned	Signed	Unsigned	Signed
1	4.10	4.25	4.04	4.08	4.30	4.32
2	3.75	4.20	3.72	4.00	4.00	4.07
3	3.79	4.05	4.16	4.15	4.07	4.11
4	3.42	3.90	3.88	4.00	3.78	4.00
5	3.57	3.95	4.20	4.23	3.96	4.04
6	3.83	4.05	4.40	4.42	4.15	4.18
7	3.35	4.00	4.16	4.31	4.11	4.25
8	3.78	4.05	4.28	4.38	4.15	4.32
9	3.52	3.65	3.60	3.35	3.33	3.54
10	4.17	4.20	4.12	4.15	3.93	4.14
11	3.08	3.55	3.04	2.96	3.30	3.32
12	3.04	3.45	3.00	3.15	3.30	3.36
13	3.87	3.90	4.36	4.38	4.22	4.29
14	3.43	3.50	3.96	4.04	4.00	4.11
15	3.52	3.85	4.12	4.08	3.44	3.86
16	3.09	3.60	3.72	4.08	3.44	3.71

The model used is

$$DIFF = a + b^{*}INSTRUCT + c^{*}LOW \qquad (1)$$

DIFF is the difference between the mean score for an item from the unsigned evaluations and the corresponding mean score from the signed evaluations. Thus, DIFF is negative for 92% of the observations. INSTRUCT equals 1 if the evaluation item pertains to the instructor, and it equals zero if the item pertains to the course. LOW equals 1 for those evaluation items where the mean score for the control group was less than 3.5, and it equals zero if the mean value for the control group was greater than 3.5. LOW was included since a greater jump in the scores might be observed for those items receiving the lowest scores; these scores have more room to increase than do the scores that are above 3.5.

A regression was first run for each of the three classes separately. Results appear in columns (1), (2), and (3) of Table 2. The signs are consistent across all regressions except for the positive constant term for section two of the professional selling class. However, this coefficient is not significantly different from zero at any reasonable level of confidence. For two of the classes, there is an overall significant increase in evaluation scores when students were asked to sign their evaluations. This is indicated by the negative constant terms that are significantly different from zero at the .06 level or better. Likewise, for these two classes when the evaluation scores are less than 3.5, the increase in evaluation scores that comes from signing the evaluations is greater. Finally, only in the second section of the professional selling class does there appear to be an effect due to whether the question pertains to the course or to the instructor. When the evaluation item pertains to the instructor, signing the evaluation results in a greater increase in the mean score than if the item concerns the course.

TABLE 2
Regression Results[a]

	(1)	(2)	(3)	(4)
CONSTANT	-.19* (-3.10)	.003 (.074)	-.082* (-2.02)	-.081 (-1.58)
LOW	-.21* (-2.49)	-.041 (-.41)	-.102** (-1.77)	-.132* (-2.87)
INSTRUCT	-.082 (-.96)	-.16* (-2.41)	-.050 (-.892)	-.097* (-2.32)
SALES1				-.206* (-4.18)
SALESM				.051 (1.05)
R^2	.35	.31	.20	.47
n	16	16	16	48

a: T-statistics are in parentheses
*: Indicates significance at the .06 level
**: Indicates significance at the .10 level

To increase the degrees of freedom, the data from the three classes were pooled, and two dummy variables were included to check for differences between classes as suggested from

the initial regression results. SALES1 equals 1 if the observation is from the first section of the professional selling course; otherwise, it is a zero. SALESM is a 1 if the observation is from the sales management course; otherwise, it equals zero. Results are in column (4) of Table 2. Again, the constant term is negative, indicating an overall increase in mean evaluation scores when the evaluations are signed. However, it is only significant at about the .15 level. Second, the coefficient on LOW is negative and significantly different from zero, indicating that signing the evaluations brought a greater increase in mean scores for those items with lower unsigned scores. The coefficient on INSTRUCT is also negative and significantly different from zero, indicating a greater increase in mean score for items pertaining to the instructor. Finally, the difference in signed and unsigned evaluation scores was greater for section one of the professional selling course.

The R^2 of the pooled regression is .47, indicating that only 47% of the difference in the sets of evaluation scores is explained by the model. This may indicate that some variables have been left out of the model. On the other hand, this low R^2 might simply indicate a large degree of random variance in the difference between the scores. If this is the case, the reliability of the evaluation is called into question.

This pilot study, though certainly not without its faults, points to problems of both validity and reliability in the evaluations administered to students. Although the motivations behind the difference in the signed scores and unsigned scores cannot be explained here, the regressions do show systematic differences in scores. In addition, the increase in scores from signing the evaluations is related to the type of evaluation asked for (course or instructor) and the level of the initial unsigned score. The low R^2 also suggests random variations in the differences between scores that cannot be explained by the regression model.

CONCLUSIONS AND RECOMMENDATIONS

This pilot study combined with the extensive literature on previous studies reveals major concerns regarding the use of student evaluations. First, the results suggest that the validity and reliability of student evaluations are questionable in terms of measuring teacher effectiveness. What does appear to be measured more accurately is the personality of an instructor. This would suggest that students are more focused upon "short-term wants" than upon "long-term needs," i.e., they evaluate their current "experience" in a particular course rather than what they "need to learn" for future job success. This is not abnormal, particularly for the majority of traditional students (which most of us encounter) as they have a very limited understanding of requirements for successful job performance. It remains the responsibility of professors to address "needs" rather than "wants."

The authors suggest, therefore, that the problem is not in the area of the evaluation instrument nor with the instructors being evaluated. The current practice of having students evaluate does measure a professor's skill in delivering the "learning process" as an interesting and enjoyable experience. That may be acceptable to those that subscribe to the philosophy that education is a process of marketing.

On the other hand, if there is interest in measuring "teaching effectiveness," academicians must focus upon what has been learned and its relationship to the work world. To do so, measurement must be obtained from relevant populations that are in a position to measure teaching effectiveness with validity and reliability. The authors suggest that such populations include former students now in the work place and the employers of those students. Surveying students that have been working for a year or more can be meaningful as they possess sufficient experience to which they can relate previous learning. Similarly, immediate supervisors of those former students can evaluate specific job performance and any accompanying strengths and/or weaknesses that have been observed.

While these additional sources of evaluation are more cumbersome to obtain, they are a necessity if teaching effectiveness is to be measured with a degree of accuracy. The present practice of using "in-house" students appears to meet only the criteria of "ease and convenience"--criteria that is inadequate when it may significantly impact a professor's career.

REFERENCES

Due to space limitations for printing and the authors' desire to retain the entire body of this article, the 29 references have not been included in this printing but are available upon request.