Evaluating Student Evaluations: A Preliminary Look at a Commonly Used Tool

Thomas I. Wahl and Kenneth Cassavant

Abstract

This paper examines the use of student evaluations of teaching (SETs) to provide teacher improvement information. The results suggest that standard deviations or other distribution information offer useful information for helping instructor preparation. It appears, based upon this preliminary case study, that SETs must be administered carefully, reported fully, and analyzed completely.

Student evaluations of teaching (SETs) are commonly used and conventionally accepted methods of instructor evaluation. Recent studies by Broder and Taylor (1994) and Casavant and Worley (1994) have investigated the use of SETs to encourage improvements in teaching and to measure student response to the change. However, SETs are usually administered at the end of the teaching period, which does not give the instructor the opportunity to make "mid-course" adjustments in response to SETs. Using multiple SETs during a term allows mid-course adjustments to be evaluated by the same set of students and the dynamics of evaluations over the entire course to be investigated.

Introduction

In this effort, multiple SETs over a term were used in a case study approach to investigate what information was conveyed and how the evaluations changed or differed over the term, over students, and over the components of the SETs, including course structure and instructor performance. Specifically, the relevance of average, and the variability of, valuation coefficients of SETs over the term by student are examined to better understand the dynamic process that leads to end of term SETs.

Case Study Approach

The class used as the case study was a new Master's level Agricultural Marketing course, which included a mix of 15 Ph.D. and Masters students. The agricultural marketing and economics background of the students ranged from very limited to Ph.D. level training. The class met for two lecture sessions each week of the 15-week semester. The SET form used included 10 questions (Table 1), which can be divided into two groups: those that relate to the course structure and content, and those that relate to instructor performance.

Students responded to the questions by circling A to F which, respectively, represented excellent, very good, good, poor, and very poor. The responses were assigned a numerical value, where A = 5, B = 4, C = 3, D = 2, E = 1, F = 0. Confidentiality of students was maintained while allowing student responses to be tracked over the term by having stu-

Table 1 Means and Standard Deviations of the Evaluation Results by Question

		Evaluation #			
Question:		#1	#2	#3	#4
Course Structure and Content:					
1	Organization of the course	3.667 (0.869)	3.400 (1.020)	3.250 (1.233)	2.933 (1.569)
2	Grading	3.500 (0.500)	3.846 (0.769)	3.857 (0.915)	3.333 (1.247)
3	General Rating of the course	3.500 (0.500)	3.467 (0.806)	3.357 (1.231)	2.933 (1.769)
4	Learning Emphasis: Independent thinking vs rote memorization	3.667 (0.850)	3.533 (0.957)	3.500 (1.041)	3.733 (1.289)
5	Examinations	2.333 (0.943)	4.500 (1.491)	3.167 (1.518)	3.133 (1.500)
Instructor Performance:					
6	Instructor Interest	4.000 (0.632)	3.667 (0.869)	3.787 (0.939)	3.467 (1.147)
7	Preparation for class	4.267 (0.573)	3.800 (0.833)	3.714 (1.030)	3.600 (0.952)
8	Presentation of subject matter	3.800 (1.046)	3.214 (0.939)	3.462 (1.151)	3.000 (1.461)
9	Attitude Toward Students	4.400 (0.712)	3.933 (0.772)	3.923 (0.828)	3.933 (0.680)
10	General Rating of the Instructor	3.923 (0.615)	3.429 (0.728)	3.462 (1.082)	3.357 (1.231)

Wahl is an assistant professor and Cassavant is a professor in the Department of Agricultural Economics, Washington State University, Pullman, Washington.

[1]The 15 students drew from a hat of 20 numbers. The last student to draw destroyed the remaining numbers.

dents draw a number from a hat and then identifying their SET by that number over the term.[1]

The SETs were administered four times over the semester. The first was at the end of the third week of the semester and prior to any homework, exams, or quizzes. Due to the diverse background of the class, the material covered during this evaluation period and each subsequent period was new to a subset of the class while simultaneously being a review for others. The second SET was administered during the sixth week of the semester prior to the mid-term exam, but subsequent to the return of graded home work. The third SET was administered following the return of the mid-term exam. The fourth and last SET was administered to students immediately following the final exam.

The timing of the SETs may influence the results of the SET in that questions about examinations prior to the mid-term may reflect perceptions if answered and skew the results. Administering the SET immediately following exams may bias the results downward if the examination was viewed as "unfair" or too difficult.

Results by Question

The means and standard deviation of the results by question for the four SETs are presented in Table 1. Traditionally, the means, and perhaps the distribution, of the fourth SET are the only evaluation results that will be available to an instructor.

The results of the fourth SET suggest there were some organizational and content problems with the course, something not entirely unexpected in the development of a new course. The instructor performance scores, while not stellar, indicate a "good" to "very good" evaluation with presentation of the subject matter being the lowest, but still being evaluated as "good".

The first SET results indicate that organization was ranked among the highest of the course structure and content questions. Presentation, however, was ranked the lowest of the instructor performance questions. Nevertheless, the mean score for presentation is still higher than any of the course structure and content questions and was nearly at the very good level. The conclusion reasonably drawn after the first SET, particularly since no grading or examination had occurred, was that overall the course was perceived to be headed in the right direction and that students seemed reasonably satisfied with the new course and its expected direction.

The results of the second SET show that 8 of the 10 means declined slightly and standard deviations increased somewhat suggesting that, in retrospect, the students were less happy with the course and there was less consensus. The slight downward trend continued in the third SET and the variance continued to increase. The fourth SET, in general, continued the slight downward trend, with increasing variance. The increasing variability of the responses suggests a growing lack of consensus about the course and perhaps diverging individual views, which may reflect the mix of Ph.D., and Masters students.

Results by Students

The tracking of individual responses over the four SETs allows the results by student to be examined. The mean and standard deviation by student and the overall mean for the four SETs are presented in Table 2. The results by student are consistent with the by-question results in that the means declined slightly from period to period. The mean decreased for 11 students, increased for 3, and stayed the same for one over the SETs. The variance decreased for 8 students, increased for 6, and stayed the same for one. Student 4 either had little doubt about his/her response, was intimidated, or was non participating in a favorable manner (at least from an instructor's point of view). The responses by student are less variable, in general, than those by question.

Difference Between the Means

The results by-question, by-student, and overall suggest a slight decline over the semester in the means of the responses.

Table 2 Means and Standard Deviations of the Results by Student Number

	Evaluation #			
Student Number:	#1	#2	#3	#4
1	3.250	3.444	3.700	4.000
	(0.433)	(0.497)	(0.458)	(0.000)
2	3.222	2.222	1.667	2.500
	(1.030)	(0.786)	(0.745)	(1.565)
3	4.600	4.143	5.000	5.000
	(0.490)	(0.350)	(0.000)	(0.000)
4	5.000	5.000	5.000	5.000
	(0.000)	(0.000)	(0.000)	(0.000)
5	3.625	3.500	2.200	2.100
	(0.696)	(0.806)	(0.600)	(1.300)
6	4.143	3.556	3.800	3.300
	(0.639)	(0.685)	(0.400)	(0.781)
7	4.500	3.667	4.000	3.600
	(0.500)	(0.667)	(0.447)	(0.490)
9	3.200	3.111	2.600	2.000
	(0.600)	(0.567)	(0.663)	(0.775)
12	3.556	3.222	2.556	1.200
	(1.066)	(0.416)	(0.831)	(0.748)
13	4.167	3.000	4.900	4.800
	(0.373)	(0.000)	(0.300)	(0.400)
14	4.000	4.000	3.111	2.100
	(0.707)	(0.707)	0.567)	(0.700)
15	4.667	4.600	4.600	4.500
	(0.471)	(0.490)	(0.490)	(0.500)
18	3.750	3.222	2.900	3.200
	(0.433)	(0.786)	(0.700)	(0.400)
19	3.300	3.111	3.300	2.889
	(0.458)	(0.314)	(0.458)	(0.314)
20	3.857	4.222	na	3.900
	(0.350)	(0.629)	na	(0.300)

A statistical test of the differences in the mean was performed on all six possible comparisons of the results by SET period (1 to 2, 1 to 3, 1 to 4, 2 to 3, 2 to 4, and 3 to 4). In all cases the results by-question indicate that there is no significant difference between the means.[2]

Tests of the differences in the means by student indicate that of the 90 possible comparisons for the 15 students, only 15 of 90, involving only 6 students, are significantly different. Of these significant comparisons, 73 percent involved SET 4, the final evaluation, which was administered immediately following the final exam. Of the significant differences, 2 out of 15 were positive changes. The results suggest that of the significant differences by student, the final evaluation was a determining factor.

Rank Test

To obtain a sense of the relative importance of the differences between the course structure and content, and the instructor performance questions, as well as the individual components of the SETs, rank tests of the results were constructed (Table 3). The rank test results show that except for SET 2, instructor performance ranked higher than course structure and content, which may reflect the first time preparation of a new course. SET 2 changed primarily because the perception of examinations changed. The rank of examinations changed from 10 to 1. However, only a few students responded to this question on SET 1 and 2, probably because no examinations had yet been given.

Presentation and overall instructor evaluation decreased dramatically in SET 2, probably because during this period a review of theory was presented which was very new material for some and completely review for others. Student comments during class and after class were extremely diverse, ranging from covering the material "too fast" to "too slow."

In general, there is a high level of consistency of the rank test except for evaluation 2. Organization and the overall rating of the course were consistently ranked poorly while attitude, interest, and preparation consistently ranked high. Learning emphasis rose steadily as the class progressed.

What Did We Learn?

Over the term, student perception seemed to vary significantly. As the semester progressed, the views of students diverged, perhaps reflecting better information about the class and instructor or simply differing student perception of that information.

The mean of the responses certainly do not capture a complete representation of student perceptions. Standard deviations or other distribution information offer as much or more information than simple means for helping instructor preparation. Increasing variability indicates a growing disparity among student assessments of the class. In contrast, means simply provide a number that may not represent widely differing viewpoints. While means are useful as a "number" for administrators, they may not be as informative when used as an instructor improvement tool.

SETs work in the sense that measures of variability blindly reveal student personalities which may be useful to evaluate "outliers." The combination of Ph.D. and Masters students in one course, with dramatic differences in theoretical background is problematic. The differing results between SET 1 and SET 2 may reflect differing responses by theory background and reinforces the importance of reporting the variability in student responses. Finally, the use of SETs over the semester revealed the dynamics in the evaluation process, reflected changing student perceptions, and reflected the flow and dynamics of classroom activities over the semester. It does appear, based upon this preliminary case study, that SETs must be administered carefully, reported fully, and analyzed completely.

Table 3 Rank Test of the SET Components

Question:		Evaluation # #1	#2	#3	#4
Course Structure and Content:					
1	Organization of the course	6	9	9	10
2	Grading	9	3	2	6
3	General Rating of the course	8	7	8	9
4	Learning Emphasis: Independent thinking vs rote memorization	7	6	5	2
5	Examinations	10	1	10	7
	Total Group Rank	40	26	34	34
	Group Average	8	5.2	6.8	6.8
Instructor Performance:					
6	Instructor Interest	3	5	3	4
7	Preparation for class	2	4	4	3
8	Presentation of subject matter	5	10	6	8
9	Attitude Toward Students	1	2	1	1
10	General Rating of the Instructor	4	8	7	5
	Total Group Rank	15	29	21	21
	Group Average	3	5.8	4.2	4.2

References

Broder, Josef M. and William J. Taylor. "Teaching Evaluation in Agricultural Economics and Related Departments." *AJAE*(76) (February 1994) 153-162.

Worley, Thomas and Kenneth Cassavant. "Student Evaluation of Teaching: A Tool for Directing and Measuring Course Improvement Efforts." NACTA, 1994.

[2]The differences in the mean test used was:

$$\mu_i - \mu_j = \frac{\bar{x}_i - \bar{x}_j}{(s_i^2/n_i + s_j^2/n_j)^{1/2}}$$