Student Evaluations of Teaching: Multiple Class Sections Compared

Meredith Uttley Lander University 320 Stanley Avenue Greenwood, SC 29649
Linda Ann Carson Lander University 320 Stanley Avenue Greenwood, SC 29649

ABSTRACT

Student evaluations of their professors and courses, originally designed to help instructors improve the quality of their instruction, are increasingly used in tenure and promotion decisions. The goal of this research is to identify external factors that influence these student evaluations. The three measures of teaching, progress on relevant course objectives, excellence of instructor, and excellence of course, do not follow a consistent pattern. Greater progress is associated with non-introductory classes and instructors teaching no more than two sections of a class. Excellence of instructor is not significantly associated with any of the external variables we analyzed. Excellence of course is associated with larger classes, shorter classes, and lower student response rates. These external factors are beyond the control of instructors, yet they live with the results.

INTRODUCTION

The goal of teaching is to educate and not surprisingly, assessing that education has become a major focus in academia. Accreditation guidelines have placed pressure on colleges and universities to place greater emphasis on teaching (Garcia and Floyd 2002; Read et al. 2001). Consequently, faculty are increasingly held responsible for the learning of their students (Garcia and Floyd 2002). The assessment method has been to have students evaluate their professors. Student evaluation of teaching (SET) has a long history, dating back to the 1920s at the University of Washington (Seldin 1993). Since the 1960's, SETs have been used in the United States, Australia, Canada, Europe and Great Britain. At that time, such evaluations were "almost entirely voluntary" (Algozzine et al. 2004:134). Such ratings have become more and more common. Seldin (1993) noted that the use of student ratings among a group of 600 liberal arts colleges increased from 29% to 68% to 86% in 1973, 1983, and 1993 respectively. Haskell (1997) found that 95% of business school deans at 220 accredited undergraduate schools use SETs as a source of information. Today, student evaluation of faculty is a mandatory part of post-secondary teaching in America (Algozzine et al. 2004). SETs are the fastest growing method of evaluation (Pratt, 1997).

SETs were originally designed to help instructors improve the quality of their instruction and courses, a formative function (Birnbaum 1999; Germain 2005; Haskell 1997; Rifkin 1995). SETs continue to provide formative feedback for instructors (Aultman 2006), but increasingly they are used by administrators for faculty reappointment, tenure, promotion, and pay increase recommendations, a summative function (Cashin and Downey 1992; Jackson et al. 1999; Pike 1998; Seldin 1999). Student evaluation forms may be designed by individual instructors, by departments, by colleges, or by outside agencies. Regardless of the instrument used, research, in the past, has supported the assumption that students make valid and reliable estimates of their learning (Hoyt and Perera 2000; Seldin 1993). SETs should be highly influenced by variables demonstrated to be strongly associated with effective teaching. However, research shows that a number of external factors, beyond the control of any instructor, may influence those evaluations (Johnson 2002). Following the work of Marsh and Roche (1997), construct-validity has come under increasing scrutiny.

Some detractors simply question the ability of college students, especially first-year students, to assess the quality of teaching based on limited experience, maintaining that students' perceptions are not necessarily based on fact, nor are they necessarily valid (Crumbley 1995; Sproule 2000; Wilson 1998). Others have focused on a plethora of variables that have no documented association with effective teaching. Davidovitch and Dan Soen (2006) found that instructors of mandatory courses received lower evaluations compared to instructors of electives. They also noted that instructors of natural sciences classes received lower evaluations than did instructors of humanities and social sciences classes or than instructors in the healthcare sciences. Kwan (1999) and Nerger et al. (2007) found academic discipline, in general, to be associated with differences in SETs. Burns and Ludlow (2005) as well as Davidovitch and Dan Soen (2006) found that class attendance altered SETs. Munz and Fallert (1998) showed that student mood was correlated with instructor and course ratings, as were characteristics of an instructor's personality (Clayson 1999; Nerger et al. 2007), particularly attractiveness ("hotness" based on ratemyprofessor) (Birnbaum 1998; Riniolo et al. 2006). Millea and Grimes (2002) and Davidovitch and Dan Soen (2006) found that females tended to evaluate teaching more positively. This applied to African-Americans, and students 35 and over as well (Davidovitch and Dan Soen 2006). Birnbaum (1998) and Marsh and Roche (1997) reported that students gave higher ratings to classes with less coursework, but Millea and Grimes (2002) did not find that to be the case. There is a general positive correlation between grades, either expected or actual, and student evaluations of teaching (Cohen 1981; Greenwald and Gillmore 1997; Johnson 2002; Millea and Grimes 2002; Nerger et al. 2007), but Marsh and Roche (2000) caution that good grades can come from higher motivation and greater interest in the subject matter and need not constitute bias. In fact, "workload, expected grades, and their relations with SETs were stable over 12 years" (Marsh and Roche 2000).

The effects of class size have been the focus for a number of researchers (Fernandez and Mateo 1998; Kwan 1999; Maurer et al. 2006; Shurden et al. 2005), examined recurrently since the 1920s, but have shown mixed results (UCSB 2004). Typically, small, medium, and large classes have been compared, but the criteria for establishing size categories have not been uniform. Evaluations for small classes seem to be higher in general (Shurden et al. 2005), although part of the variance may depend on the type of class (seminar, lab, lecture, or level) rather than simply the size (Nerger et al. 2007). Fernandez and Mateo (1998:599) found "a statistically significant but very small negative linear relationship between class size and mean ratings of quality of teaching". When they divided the same data into five categories, rather than three they saw a clear, U-shaped relationship between the variables (Fernandez and Mateo 1998). Results that change based on ordinal-level cut points are problematic for any research and would explain the mixed results reported in the literature.

Shurden, Smith, and Tolbert (2005) explored the effect of length of class, which corresponds to frequency of class meetings. Monday/Wednesday afternoon classes met for 75 minutes as did Tuesday/Thursday classes. Monday/Wednesday/Friday classes until 1:50 PM met for 50 minutes. While instructors' desires may be considered, class schedules are arranged to meet the requirements of university administrators. Introductory classes are spread across sections to allow the greatest number of students to enroll. Required classes must not conflict. Popular electives are scheduled in sections to attract new students to department courses. Any systematic association of SETs and length or frequency of class could dramatically affect instructors' ratings, but be totally beyond their control. Those of us who teach summer school, when classes meet four days per week have noticed that grades are typically higher. Students and instructors alike believe that this results from greater frequency of contact with the material. If so, it would be logical for classes that meet more frequently during regular terms to have higher SETs. Shurden et al. (2005) found this to be the case with significantly higher SETs for the more frequent classes.

Last year, we compared student perceptions of distance education classes taught at separate sites. As expected, we found that students from different sites held different perceptions. We thought that we understood the reasons for those disparities. Students in the room with the instructor perceived more progress on relevant objectives and rated both the instructor and the class more positively as did students who experienced fewer technical difficulties. However, we have also noticed differences in the student perceptions from different sections of classes taught by one instructor in a traditional classroom, in some cases with variation as great as we saw for the distance education sections. The difference in evaluations among an instructor's multiple sections ranges from zero to 25%. We maintain no illusion that we understand these differences. They may exist for only a few classes, or a few instructors, or for certain departments or because of external factors. In any case, these inconsistent student perceptions continue to raise concerns about the validity and reliability of SETs in general.

METHODOLOGY

The focus of this paper is to examine student perceptions in multiple sections of a class taught by a single instructor. Based on the literature, we will explore differences based on those external variables to which we have access. In 2004, Lander University adopted the Individual Development and Educational Assessment (IDEA) instrument to collect student perceptions of their classes. The growing use of and reliance on student evaluations of teaching for summative decisions and increasing questions about their construct validity, led to the development of the IDEA instrument at Kansas State University during the 1968-1969 school year (Hoyt and Cashin 1977). Now, 60,000 classes are assessed annually, providing nationally comparative data (http://www.idea.ksu.edu/StudentRatings/index.html). What distinguishes the IDEA instrument from other SETs is that student ratings are adjusted on the basis of a student's motivation to take the course, work habits and effort, as well as perceived course difficulty and class size (Hoyt and Lee 2003). Regardless of these admirable attempts to control for external influences, limitations to the interpretation of the results from this instrument remain. The IDEA system is not recommended for classes with less than 10 students due to decreased validity and reliability (Hoyt and Cashin 1977). Nearly four percent of Lander classes have fewer than 10 students enrolled. Ratings are not considered to be representative unless at least 75% of students respond (IDEA Center 2004). Only about 65% of Lander classes get a response rate of 75% and above.

The IDEA instrument provides three summary measures for each class. The first measures effective teaching in terms of progress made on particular course objectives. Instructors pick from a list of 12 possible course objectives, designating each as essential, important, or not important to the particular course. Those objectives deemed essential are double weighted in the IDEA assessment calculations. Objectives are organized into six categories: *Basic Cognitive Background; Application of Learning; Expressiveness; Intellectual Development; Lifelong Learning;* and *Team Skills* (www.idea.ksu.edu). Most categories include multiple possible objectives, rated by students on a five-point scale from *no apparent progress* (1) to *exceptional progress* (5):

Basic Cognitive Background

1. Gaining factual knowledge (terminology, classifications, methods, trends)
2. Learning fundamental principles, generalizations, or theories

Application of Learning

3. Learning to *apply* course material (to improve thinking, problem solving, and decisions)
4. Developing specific skills, competencies, and points of view needed by professionals in the field most closely related to this course

Expressiveness

6. Developing creative capacities (writing, inventing, designing, performing in art, music, drama, etc.)

8. Developing skill in expressing oneself orally or in writing

Intellectual Development

7. Gaining a broader understanding and appreciation of intellectual/cultural activity (music, science, literature, etc.)
10. Developing a clearer understanding of, and commitment to, personal values
11. Learning to *analyze* and *critically evaluate* ideas, arguments, and points of view

Lifelong Learning

9. Learning how to find and use resources for answering questions or solving problems
12. Acquiring an interest in learning more by asking questions and seeking answers

Team Skills

5. Acquiring skills in working with others as a member of a team.

The IDEA guidelines recommend, "that the meaning of the objectives [should be] discussed with your class early in the semester so a common understanding is reached". It is questionable, however, how well specific objectives are remembered when evaluations are performed later in the semester. The second measure is based on the single statement, "Overall, I rate this instructor an excellent teacher." Response options range from *definitely false* (1) to *definitely true* (5). The third measure is based on the single statement, "Overall, I rate this course as excellent." The responses are the same as those used for the second measure. For each measure, the IDEA center calculates raw and adjusted average scores, recommending that the adjusted scores be used for comparisons. Scores are adjusted on the basis of students' professed desire to take the course, expressed effort put forth, and perceived amount of work required.

We have data from 58 instructors, who taught 86 pairs of classes, 20 who taught 21 sets of three classes, two who taught four sections, three who taught five sections, and one who taught six sections for a total of 254 classes. We will compute the difference between the maximum rating and minimum rating for pairs and sets of classes taught by each specific instructor. If the IDEA instrument relies on factors correlated highly with teaching, we would expect little difference between ratings from difference sections of a class taught by a single instructor. In addition we will explore the ratings and variance in ratings between semesters, and between pairs of classes and larger sets. Further, we will explore evaluations and response rate based on length of class, class time, and class level. Our analyses will use paired t-tests, analysis of variance, and correlation (Fox 2003; Patterson and Basham 2006). Based on previous research, we have only one hypothesis, that classes that meet more frequently will have higher evaluations for all three summary measures.

ANALYSES

As with any interesting research, our exploration has generated a series of new questions for the future. All of the factors we have explored are, for the most part, beyond an instructor's control and unrelated to the process of teaching, itself. Due to the lack of literature or conflicting literature, we approached our exploration with a series of questions. Since the IDEA instrument provides three summary measures for each class, we have explored each question for each summary measure. As we found in our previous research, different measures show different results. For only some of these differences are there any suggested or even inferred explanations in the literature.

Question 1: Do SETs differ depending on whether a class is introductory level or a higher level? Introductory classes, while not limited to first-year students, certainly have a higher percentage of such students enrolled. Table 1 shows no significant difference in SETs for either excellence of instructor or excellence of course based on class level. However, figure 1 shows a highly significant (P=0.000) positive relationship between SETs for the measure, progress on relevant objectives. These results may

support the suggestion of Crumbley (1995), Sproule (2000), and Wilson (1998), who question the ability, particularly of first-year students, to assess the quality of teaching based on limited experience. It appears, however, that it is not the quality of teaching they are inexperienced in assessing. Rather, they may lack experience in assessing the progress made on the relevant objectives for the class. Students would have developed impressions of teaching and course quality in high school, but never have experienced assessing progress on course objectives. Moreover, the difference in SETs disappeared for the spring semester, further suggesting that inexperience among first-year students explained the fall disparity.

TABLE 1. Student evaluation of teaching based on class level.				
Excellence of Instructor	N	Mean Evaluation	Std. Dev.	Prob.
Introductory Classes	146	4.163	0.500	0.322
Sophomore level & above	111	4.226	0.513	
Excellence of Course				
Introductory Classes	146	3.902	0.446	0.442
Sophomore level & above	111	3.951	0.583	
Progress on Relevant Objectives				
Introductory Classes	146	3.821	0.431	**0.000**
Sophomore level & above	111	4.059	0.410	
Progress on Relevant Objectives - Fall 2005				
Introductory Classes	110	3.777	0.441	**0.000**
Sophomore level & above	54	4.131	0.381	
Progress on Relevant Objectives – Spring 2006				
Introductory Classes	36	3.953	0.378	0.675
Sophomore level & above	57	3.989	0.428	

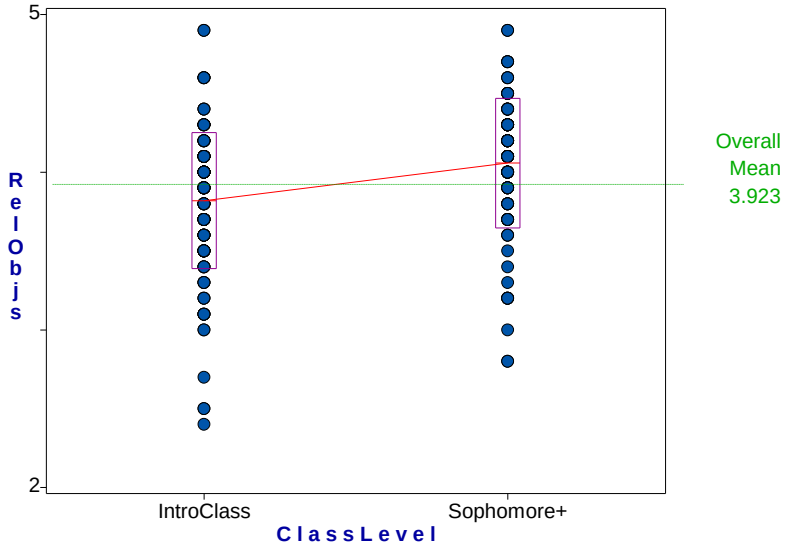

FIGURE 1: Student evaluation of progress on relevant objectives based on class level.

Question 2: Do SETs differ based on class size? The literature reports conflict for this question. Small, medium, and large classes were compared, but results have varied depending on the ordinal level cut-points used for classification (Fernandez and Mateo 1998). Instead, we correlated the number of enrolled students with the student evaluations. Table 2 shows the results of these comparisons. Class enrollment is unrelated to perceived progress on relevant objectives and excellence of instructor. In contrast, figure 2 shows that the number of students enrolled in a course is positively and significantly (P = 0.038) related to the evaluation of the excellence of the course. These results conflict with those of Shurden et al. (2005) as well as Fernandez and Mateo (1998) and are contrary to common sense, which suggests that

students can get more attention and help in a smaller class. Nerger et al. (2007) suggested that the variance in SETs may depend on the type of class (seminar, lab, lecture) rather than simply class size. It may be that class size would be more appropriately used as a relative measure, rather than an absolute, as with enrollment. At Lander, 60 students make a huge class, while at many schools, an enrollment of 60 would be considered a smaller medium size class. Given perceptual differences, however, the actual boundaries between class size categories might be university specific as well as class-type specific. If class size is influencing SETs, it would seem that the size-category boundaries should be student driven. Nothing in the literature indicates that student perceptions have been included while setting class-size boundaries, however. Trying to separate the class size and class type effects remains for the future.

TABLE 2. Student evaluation of teaching based on class enrollment.				
Summary Measure	N	Line Equation	r	Prob.
Progress on Relevant Objectives	264	Y = 3.937 + -0.000X	-0.011	0.430
Excellence of Instructor	264	Y = 4.163 + 0.001X	0.027	0.332
Excellence of Course	264	Y = 3.790 + 0.005X	0.110	**0.038**

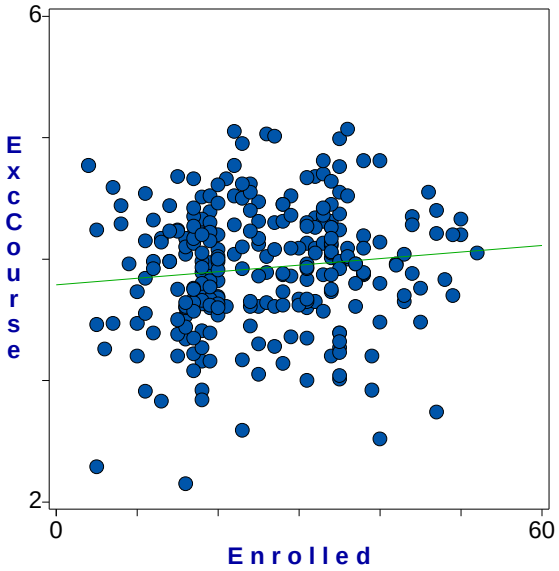

FIGURE 2: Student evaluation of excellence of course based on class enrollment.

Question 3: Do SETs differ based on the time of the class? We all know that some class times are more popular than others if students have a choice. We also know that individuals experience their peak performance at different times. Might these time preferences and peak times be related to student ratings because students want to be somewhere else or engaged in some other activity? The same items influence faculty members as well and might affect an instructor's delivery of a course, which would systematically influence SETs. For analysis, we divided class times into five categories based on student popularity. Whether these time preferences correlate with peak performance times is unexplored and unexplorable given the available data. As seen in figures 3a, b, and c, none of the relationships between SETs and class times is significant. However, student evaluations are consistently highest for the earliest time slots, typically the last populated classes when students have a choice of class times. Instructors openly talk about students sleeping in 8:00 classes or having to allow students to bring coffee to class in order to stay awake. Evaluations are consistently lowest for 10:00 and 11:00 classes, class sections that are among the first to fill. Why there would be any specific pattern between progress on relevant class objectives and class time is a mystery. Finding a pattern between the summary measures that focus on excellence of the instructor and of the course and class time seems more logical. Yet, the pattern for evaluations and excellence of the instructor (P = 0.118) nears significance along with progress on objectives (P = 0.192). In contrast, the pattern for evaluations and excellence of course is far from significant (P = 0.417).

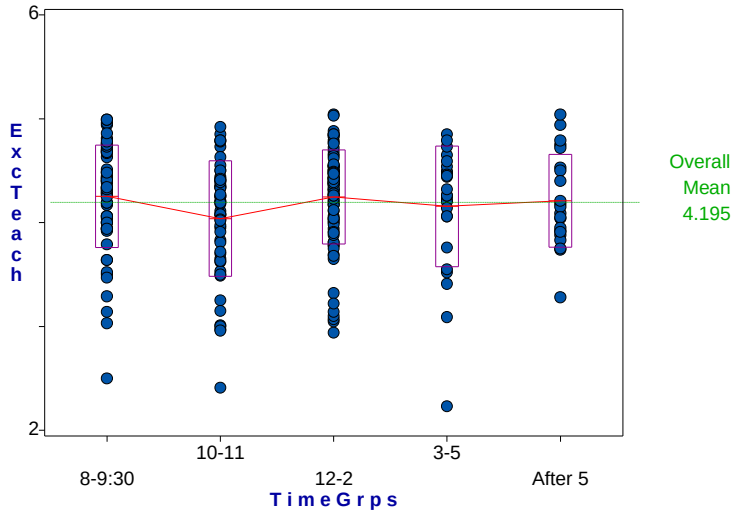

FIGURE 3a. Student evaluation of progress on relevant objectives based on class time (P = 0.192).

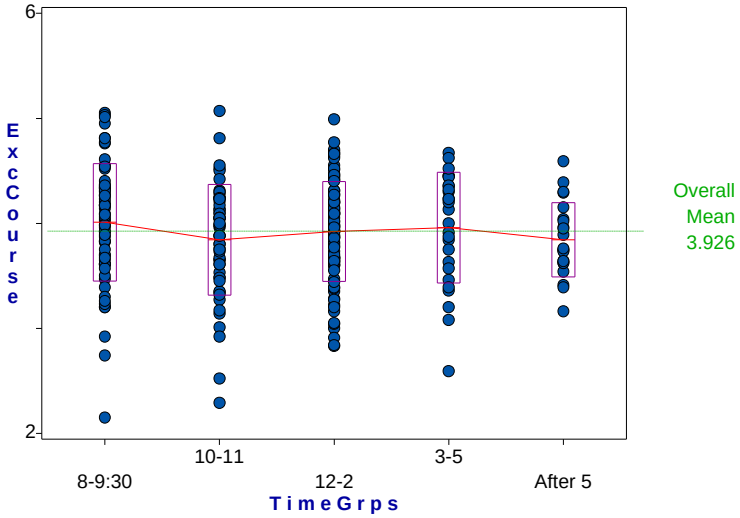

FIGURE 3b. Student evaluation of excellence of instructor based on class time (P = 0.118).

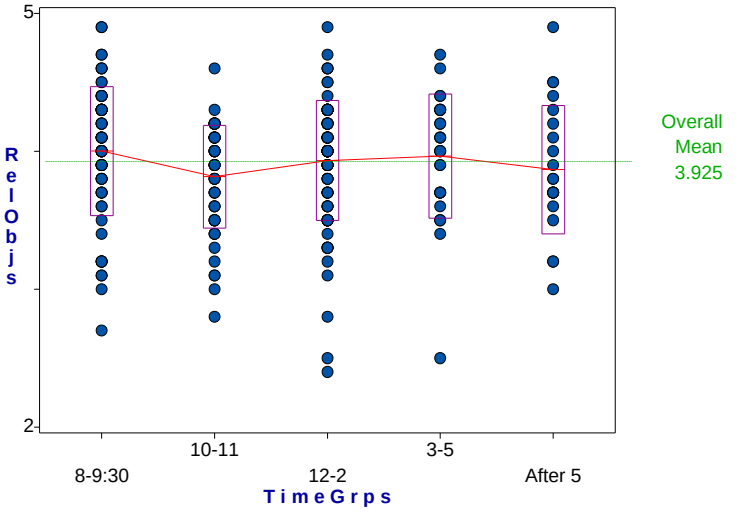

FIGURE 3c. Student evaluation of excellence of course based on class time (P = 0.417).

Question 4: Do SETs vary based on how frequently a class is taught? The frequency of class meetings is synonymous with the length of each class session. Shurden et al. (2005) found significantly higher SETs for the more frequent classes, but their research was limited to Lander's College of Business. We have expanded their research to include classes from across campus. Our hypothesis is that SETs will be higher for all three summary measures for classes that are taught more frequently. Table 3 includes the results, none of which shows significant differences in evaluations by frequency of class meeting. Nonetheless, results for both excellence of course and progress on objectives do support our hypothesis. Evaluations are highest for courses that meet three days a week. Only the evaluations for excellence of course, however, are near significance.

TABLE 3. Student evaluation of teaching based on frequency of class meetings (length of class).

Summary Measure	N	Mean Rating	Std. Dev.	Prob.
Progress on Relevant Objectives				
One Day Class	49	3.878	0.398	0.290
Two Day Class	142	3.906	0.466	
Three Day Class	71	3.992	0.398	
Excellence of Instructor				
One Day Class	49	4.257	0.426	0.240
Two Day Class	142	4.148	0.535	
Three Day Class	71	4.251	0.481	
Excellence of Course				
One Day Class	49	3.793	0.385	0.097
Two Day Class	142	3.946	0.520	
Three Day Class	71	3.990	0.545	

Question 5: How do class times, frequency and enrollment influence the response rate of SETs? The IDEA Center (2004) cautions that ratings are not considered to be representative unless at least 75% of students respond. Only about 65% of Lander classes get a response rate of 75% and above. That leaves slightly more than one third of classes that have received too few evaluations. This presents problems for instructors who rely on these evaluations for a formative function. This may also disadvantage an instructor when administrators use the evaluations for a summative function. Student evaluations of teaching are conducted during the latter part of a semester. In the College of Business and Public Affairs, evaluations for all classes are scheduled over a two-day period. So, students taking business, political science, and sociology classes could readily become aware of the pattern and avoid the evaluations, but no such regular predictable evaluation pattern exists for other departments. Figure 4 shows the slightly negative relationship between evaluation response rate and class time (measured on a 24 hour clock). While not significantly different (P = 0.214), evening classes do receive lower response rates, but the relationship is very weak (r = -0.049).

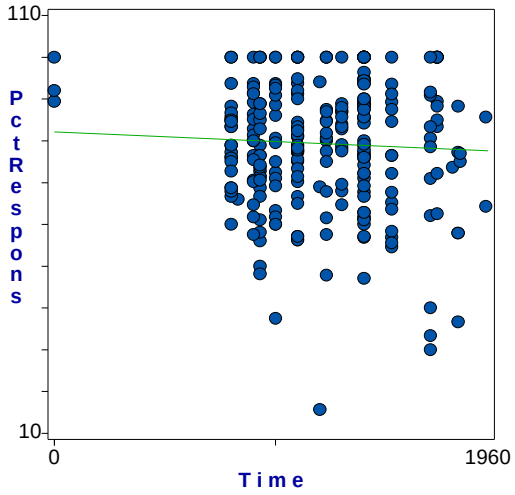

FIGURE 4. Student response rate based on class start time.

The frequency of class meetings is significantly related to the evaluation response rate (P = 0.001) with less frequent classes receiving higher response rates. These results are shown in figure 5. Lander's evening classes are taught on either one night or two nights, and while those meeting one night per week have higher response rates, evening classes as a whole, shown in figure 4 have lower response rates. Figure 6 shows the results for evaluation response rate and class enrollment. The relationship is negative, moderately weak (r = -0.237), but significant (P = 0.000). The top row of data points represents the classes that received 100% response rates, a rarity. Sophomore level classes and above receive a slightly higher response rate. The results are near, but not significant (P = 0.122).

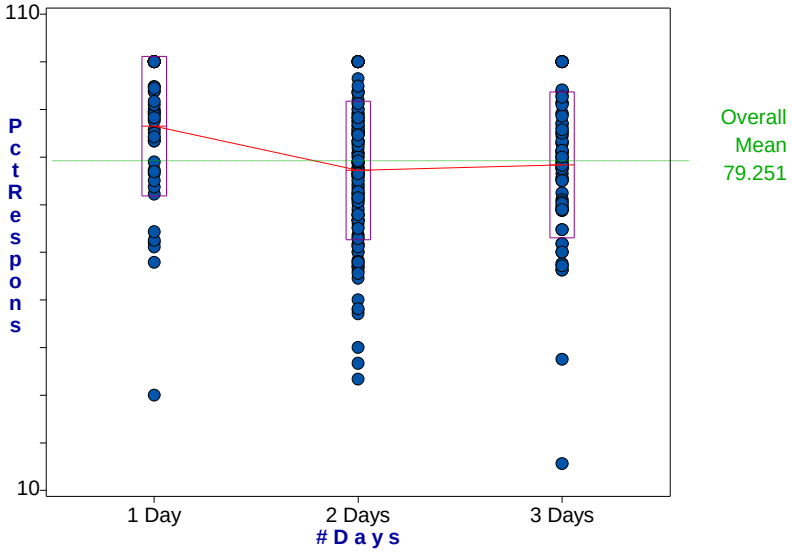

FIGURE 5. Student response rate based on class frequency.

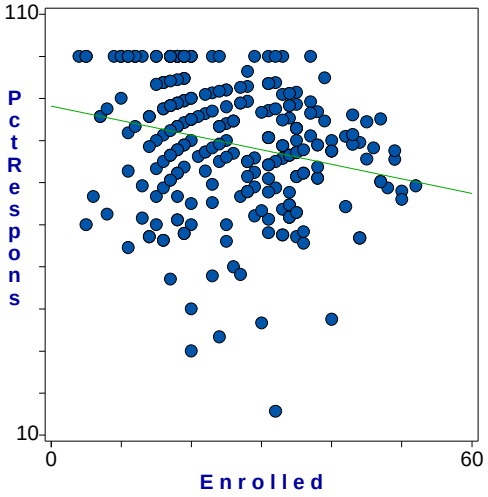

FIGURE 6. Student response rate based on class enrollment.

Question 6: The next logical question is whether evaluation response rates influence SETs? If they do, what is their influence? Table 4 shows that evaluation response rates have little effect on student evaluations of progress on relevant objectives or excellence of instructor. However, the negative effect of response rate on excellence of course is near significance. What a conundrum, a not representative set of responses may provide a higher evaluation.

TABLE 4. Student evaluation of teaching based on response rate.

Summary Measure	N	Line Equation	r	Prob.
Progress on Relevant Objectives	264	Y = 3.875 + 0.001X	0.022	0.362
Excellence of Instructor	264	Y = 4.200 + -0.000X	-0.001	0.492
Excellence of Course	264	Y = 4.137 + -0.003X	-0.079	0.102

Descriptive Statistics for Paired and Group Data: Table 5 lists the means, standard deviations, and Z-scores for the three IDEA summary measures. For each measure, there are three sets of statistics, one for the two semesters combined, one for the fall semester, and one for spring. All data distributions are positively skewed as demonstrated by the minimum and maximum Z-scores. Two patterns emerge. First, the evaluations for progress on relevant objectives express less variation than the other two measures. Second, for each measure, evaluations for the spring semester are higher than for the fall; spring evaluations also express consistently greater variance.

TABLE 5. Student evaluation of teaching descriptive statistics.

Summary Measure	N	Mean Evaluation	Std. Dev.	Min Z-score	Max Z-score
Progress on Objectives	109	0.345	0.244	-1.412	3.499
Fall 2005	70	0.333	0.236	-1.412	2.402
Spring 2006	39	0.367	0.261	-1.405	3.194
Excellence of Instructor	109	0.367	0.311	-1.180	4.612
Fall 2005	70	0.334	0.273	-1.222	3.430
Spring 2006	39	0.426	0.366	-1.110	3.760
Excellence of Course	109	0.429	0.369	-1.163	4.283
Fall 2005	70	0.420	0.324	-1.297	3.056
Spring 2006	39	0.445	0.443	-0.983	3.532

Differences among multiple sections: We have data for 86 pairs of classes and for 27 sets of three or more sections of the same class by the same instructor. Are the evaluations for these two categories of data equivalent? Is the variation in evaluation scores for the paired classes comparable to the variation for groups of classes? Table 6 lists the mean evaluations, standard deviations, Z-scores, and significance comparing evaluations for classes by pairs and groups. The within group variation is consistently higher for groups of classes than for pairs of classes. Data for progress on relevant objectives for pairs of classes is almost normally distributed. The remaining groups of data are negatively distributed, particularly the data for excellence of instructor, the evaluations for which are virtually indistinguishable. The difference in evaluations for excellence of course is near significant (P = 0.171), with instructors teaching only two sections of a class faring slightly better. The only difference that is significant is for the comparison of progress on relevant objectives (P = 0.006). Students thought that they made more progress when an instructor taught just two sections of a class.

TABLE 6. Descriptive statistics for SETs: pairs *versus* groups of classes.

Summary Measure	N	Mean Evaluation	Std. Dev.	P	Min Z-score	Max Z-score
Progress on Objectives						
Paired data	172	3.978	0.408	**0.006**	-2.885	2.256
Grouped data	92	3.825	0.468		-3.046	2.298
Excellence of Instructor						
Paired data	172	4.196	0.481	0.980	-3.711	1.754
Grouped data	92	4.198	0.538		-3.657	1.380
Excellence of Course						
Paired data	172	3.960	0.496	0.171	-3.368	2.199
Grouped data	92	3.870	0.523		-3.288	2.294

Table 7 shows that the differences between the minimum and maximum evaluations based on instructors are always significantly higher when they teach three or more sections of a class rather than just two sections.

TABLE 7. Within group variation: data for pairs *versus* three or more sections of a class.				
Summary Measure	N	Mean Difference	Std. Dev.	Prob.
Progress on Relevant Objectives				
Paired data	82	0.313	0.239	**0.018**
Grouped data	27	0.441	0.239	
Excellence of Instructor				
Paired data	82	0.331	0.317	**0.034**
Grouped data	27	0.476	0.267	
Excellence of Course				
Paired data	82	0.376	0.369	**0.008**
Grouped data	27	0.591	0.326	

Question 7: Our last question is whether SETs are significantly different for two sections of a class taught by the same instructor. These analyses are limited to data for pairs of classes. Table 8 shows the mean evaluations, variance, correlation, *t* statistic and probability. We analyzed data for the combined semesters, then fall 2005, and spring 2006 separately. In fall 2005, differences between an instructor's two class sections were significant for both progress on relevant objectives (P = 0.05) and excellence of instructor (P = 0.02). The differences for excellence of course were significant for the year, but not for either semester alone. For some pairs of classes, the variance among evaluations is as little as 0.01; for some it more than doubles from one class to the other (0.16 compared to 0.36).

TABLE 8. Student evaluation of teaching for pairs of classes.								
Summary Measure	N	Mean Evaluations		Variance		r	2 tail *t*-Stat	Prob.
		Class1	Class2	Class1	Class2			
Progress on Objectives	86	3.96	3.97	0.16	0.18	0.55	-0.86	0.39
Fall 2005	51	3.90	4.00	0.18	0.17	0.63	-2.04	**0.05**
Spring 2006	35	4.05	3.99	0.10	0.19	0.48	0.92	0.36
Excellence of Instructor	86	4.17	4.22	0.22	0.24	0.56	-0.98	0.33
Fall 2005	51	4.11	4.23	0.25	0.21	0.73	-2.45	**0.02**
Spring 2006	35	4.26	4.19	0.17	0.28	0.36	0.68	0.50
Excellence of Course	86	3.83	4.02	0.24	0.22	0.62	-3.27	**0.00**
Fall 2005	51	4.05	3.98	0.16	0.36	0.35	0.66	0.52
Spring 2006	35	3.92	4.00	0.22	0.27	0.47	-1.56	0.12

DISCUSSION

Our analyses have revealed some significant differences among the SETs, but overall, confusing and conflicting results. Absolutely clear is that the three measures of teaching do not follow a consistent pattern. Table 9 summarizes our findings, showing the influence of each external variable on each measure of teaching. Larger classes, classes that meet more frequently, and classes with lower student response rates show significant or near significant relationships with excellence of course (P = 0.038, 0.097, and 0.102 respectively), the measure most influenced by external factors. Evaluation response

rate, itself, is associated with other external factors. They are higher for classes that meet less frequently (P = 0.001) and for smaller classes (P = 0.000) suggesting that external factors influence evaluations both directly and indirectly, creating a very complicated relationship. Evening classes in general receive lower evaluations, but those sections meeting one night per week have higher response rates and receive higher evaluations for excellence of course.

Students in classes designed to be taken by sophomores or higher level students think they have made more progress on relevant objectives for the year (P = 0.000), particularly in fall 2005 (P = 0.000), compared to students in introductory classes. This finding supports suggestions that the lack of evaluation experience among beginning college students affects SETs. Class type, whether pairs or groups, is significantly related only to progress on relevant objectives for the year (P = 0.006), as opposed to single semesters, with instructors teaching two sections receiving higher evaluations.

TABLE 9. Summary of the influence of external variables on the IDEA measures of teaching.

External Influence	*Progress on Objectives*			*Excellence of Instructor*			*Excellence of Course*		
	Year	Fall	Spring	Year	Fall	Spring	Year	Fall	Spring
Class level (intro/soph+)	**0.000**	**0.000**	n.s	n.s.	--	--	n.s.	--	--
Class enrollment	n.s.	--	--	n.s.	--	--	**0.038**	--	--
Class time	n.s.	--	--	n.s.	--	--	n.s.	--	--
Class meeting frequency	n.s.	--	--	n.s.	--	--	0.097	--	--
Response rate	n.s.	--	--	n.s.	--	--	0.102	--	--
Pairs *vs* groups	**0.006**	--	--	n.s.	--	--	n.s.	--	--
Class 1 *vs* class 2	n.s.	**0.05**	n.s.	n.s.	**0.02**	n.s.	**0.00**	n.s.	n.s.

Excellence of instructor does not demonstrate a significant relationship with any of the external variables that we analyzed. This allows the possibility that this variable does specifically measure quality of teaching. Evaluations may in fact, be less questionable in terms of validity if this single measure were the output. Progress on relevant objectives is significantly related to whether or not a class is introductory or higher level and whether an instructor teaches two or more than two sections, with those teaching two sections or higher level classes receiving slightly higher evaluations.

If instructors wanted to maximize their evaluations, they would teach only two sections of a class in one semester. They would teach classes that meet one night per week or early on Monday, Wednesday, and Friday mornings. They would never teach at 10:00 AM or 11:00 AM, or classes that meet twice per week in the evening. They would teach upper level classes that are relatively large and they would hope for a relatively low response rate. SETs were originally designed to help instructors improve their teaching and their classes, but when a number of external factors influence those evaluations, how do instructors decide what changes to make? Should they average their evaluation scores over a school year or over several semesters of teaching a specific class? What if a class is small one semester, but large another or if one section is small, but another section large? What if the class meets at 8:00 AM one semester, but at 10:00 AM another semester? How are instructors to make sense of these evaluations? It is scary to realize that administrators are increasingly relying on these SETs to make decisions about an instructor's value, tenure, promotion, and pay raises.

REFERENCES

[1] Algozzine, B., J. Beattie, M. Bray, C. Flowers, J. Gretes, L. Howley, G. Mohanty, and F. Spooner. 2004. Student evaluation of college teaching: a practice in search of principles. *College Teaching* 52.4:134(8).

[2] Aultman, L. 2006. An unexpected benefit of formative student evaluations.(QUICK FIX). *College Teaching* 54.3:251.

[3] Birnbaum, M.1999. A survey of faculty opinions concerning student evaluations of teaching. *The Senate Forum: A publication of the Academic Senate of California State University, Fullerton, 14*:19-22.

[4] Burns, S., and L. Ludlow. 2005. Understanding Student Evaluations of Teaching Quality: The Contributions of Class Attendance. *Journal of Personnel Evaluation in Education* 18(2):127-138.

[5] Cashin, W., and R. Downey. 1992. Using global student rating items for summative evaluation. *Journal of Educational Psychology* 84:563-572.

[6] Clayson, D. 1999. Students= evaluation of teaching effectiveness: Some implications of stability. *Journal of Marketing Education* 21:68-75.

[7] Cohen, P. 1981. Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research* 51:281-308.

[8] Crumbley, L. 1995. The dysfunctional atmosphere of higher education: Games professors play, *Accounting Perspectives* 1(1)

[9] Davidovitch, N. and D Soen. 2006. Class attendance and students' evaluation of their college instructors. *College Student Journal* 40.3:691(13).

[10] Fernandez, J., and M. Mateo. 1998. Is there a relationship between class size and student ratings of teaching quality? *Educational and Psychological Measurement* 58:596(9).

[11] Fox, William. 2003. *Social Statistics*, fourth edition. Belmont, California: Wadsworth Group.

[12] Garcia, J. and C. Floyd. 2002Addressing evaluative standards related to program assessment: how do we respond? *Journal of Social Work Education* 38.3:369(14).

[13] Germain, M. and T. Scandura. 2005. Grade inflation and student individual differences as systematic bias in faculty evaluations. *Journal of Instructional Psychology* 32.1:58(10).

[14] Greenwald, A.., and G. Gillmore. 1997. Grading leniency is a removable contaminant of student ratings. *American Psychologist* 52:1209-1217.

[15] Haskell, R. 1997. Academic Freedom, Promotion, Reappointment, Tenure and The Administrative Use of Student Evaluation of Faculty: (Part II) Views From the Court. *Education Policy Analysis Archives* 5(17).

[16] Hoyt, D. and W. Cashin. 1977. IDEA TECHNICAL REPORT NO. 1: Development of the IDEA System. *IDEA Technical Report #1*. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

[17] Hoyt, D. and E. Lee. 2003. Understanding The IDEA System's Extraneous Variables. *IDEA Research Report #6*. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

[18] Hoyt, D. and S. Perera. 2000. Validity of the IDEA Student Ratings of Instruction System: An Update. *IDEA Research Report #2*. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

[19] IDEA Center. 2007. www.idea.ksu.edu/StudentRatings/index.html, accessed May 25.

[20] IDEA Center. 2004. Interpretive Guide: IDEA Diagnostic Form Report. www.idea.ksu.edu/diagnosticguide.pdf

[21] IDEA Center. www.idea.ksu.edu.

[22] Jackson, D., C. Teal, S. Raines, T. Nansel, R. Force, and C. Burdsal. 1999. The Dimensions of Students' Perceptions of Teaching Effectiveness. *Educational and Psychological Measurement* 59:580-596.

[23] Johnson, Valen. 2002. Teacher Course Evaluations and Student Grades: An Academic Tango. *Chance* 15:9-16.
[24] Kwan, K. 1999. How fair are student ratings in assessing the teaching performance of university teachers? *Assessment of Evaluation in Higher Education* 24:181-195.
[25] Marsh, H., and L. Roche. 2000. Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology* 92:202-228.
[26] Marsh, H., and L. Roche. 1997. Making students' evaluations of teaching effectiveness effective: The central issues of validity, bias, and utility. *American Psychologist* 52:1187-1197.
[27] Maurer, T., J. Beasley, J. Dilworth, A. Hall, J. Kropp, M. Rouse-Arnett, and J. Taulbee. 2006. Child and family development students polled: study examines student course evaluations. *Journal of Family and Consumer Sciences* 98.2:39-45.
[28] Millea, M., and E. Grimes. 2002. Grade expectations and student evaluation of teaching. *College Student Journal* 36:582-590.
[29] Munz, D., and A. Fallert. 1998. Does classroom setting moderate the relationship between student mood and teaching evaluations? *Journal of Social Behavior & Personality* 13(1):23-32.
[30] Nerger, J., W. Viney, and R. Riedel II. 1997. Student ratings of teacher effectiveness: use and misuse. *The Midwest Quarterly* 38:218(16).
[31] Patterson, D.A., and R.E. Basham. 2006. *Data analysis with spreadsheets.* Pearson Allyn and Bacon.
[32] Pike, C. 1998. A validation study of an instrument designed to measure teaching effectiveness. *Journal of Social Work Education* 34:261-271.
[33] Pratt, D. 1997. Reconceptualizing the evaluation of teaching in higher education. *Higher Education* 34:23-44.
[34] Read, W., D. Rama and K. Raghunandan. 2001. The relationship between student evaluations of teaching and faculty evaluations. *Journal of Education for Business* 76:189-192.
[35] Rifkin, T. 1995. Eric Review: Faculty Evaluation in Community Colleges. *Community College Review* 01061995:63-72.
[36] Riniolo, T., K. Johnson, T. Sherman and J. Misso. 2006. Hot or not: do professors perceived as physically attractive receive higher student evaluations? *The Journal of General Psychology* 133.1:19(17).
[37] Seldin, P. 1993. The Use and Abuse of Student Ratings of Professors. *The Chronicle of Higher Education* 39:40-42, July 21.
[38] Seldin, P. 1999. Current practices—good and bad—nationally. In P. Seldin (Ed.), *Changing Practices in Evaluating Teaching* (pp.1-24). Bolton, MA: Anker Publishing Company.
[39] Shurden, M., S. Smith, and S. Tolbert. 2005. Faculty Evaluations: The Effects of Class Size and Length of Class Period, presented at conference, abstract only, *Proceedings: Ciber Institute's Teachers and Learning Conference.*
[40] Sproule, R. 2000. Student evaluation of teaching: A methodological critique of evaluation practices.Education Policy Analysis Archives 8(50), http://epaa.asu.edu/epaa/v8n50.html.
[41] UCSB. 2004. Teaching large classes. Office of Academic Programs, Instructional Development University of California, Santa Barbara. *http://www.oic.id.ucsb.edu/Resources/Teaching/Large.ucsb.html*
[42] Wilson, R. 1998. New research casts doubt on value of student evaluations of professors. *The Chronicle of Higher Education* January:A12-A14.