

A Brief Analysis of the Student Evaluation of Teaching Effectiveness

Ross Moran, Office of Institutional Research
March, 2001

This is a review of some of the statistical properties of the *Student Evaluation of Teaching Effectiveness* (SETE) course evaluation form used at CSUSB. The first part examines the psychometric characteristics of the instrument, the second considers several normative groupings, and the third reports on a preliminary multivariate analysis as to which variables predict students' rating of a course. The paper concludes with some recommendations.

Data used for this study were the 58,234 ratings of 2,951 class sections "SETE'd" in Winter, Spring and Fall 2000. A copy of the SETE instrument is attached. The form consists of 10 multiple response items which purport to address various aspects of the instructional experience, each item is responded to using a four point scale consisting of the descriptors "Excellent," "Good," "Fair," and "Poor." The reverse side of the form includes five items soliciting comments on the instructor's organization, approach to grading, availability for consultation, teaching effectiveness and general comments. This research dealt only with the numeric items.

It has been widely observed that the responses skew with most ratings at the positive side of the scale. The following table lists the ten items and the frequency of selection for each alternative.

Item	4 pt. Avg.	Percent of Valid Ratings				Pct. Omit
		Ex'lent	Good	Fair	Poor	
1	3.76	79.3	18.0	2.4	0.3	0.1
2	3.63	69.7	24.5	4.9	0.9	0.1
3	3.50	61.7	28.5	7.9	1.8	0.3
4	3.60	69.9	22.1	6.4	1.7	0.1
5	3.43	58.2	28.9	10.2	2.7	0.3
6	3.46	59.9	28.4	9.1	2.5	0.6
7	3.49	61.2	28.5	8.3	2.0	0.3
8	3.49	62.7	26.1	8.6	2.5	0.2
9	3.48	62.5	25.7	9.0	2.8	0.2
10	3.53	63.6	27.3	7.4	1.7	0.1
Avg.	3.54	64.9	25.8	7.4	1.9	0.2

The Numeric SETE Items:

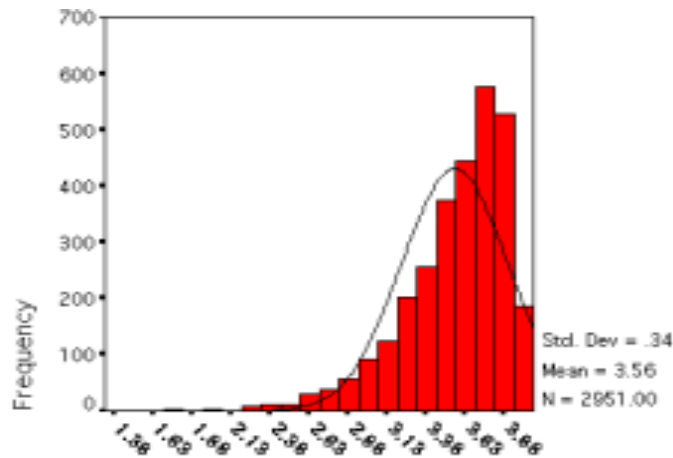
1. Rate how well your instructor knew the subject matter of the course.
2. Rate how well prepared your instructor was for class sessions.
3. Rate how well your instructor organized the course.
4. Consistent with class size, rate how well your instructor encouraged and was responsive to student questions and comments.
5. Consistent with class size, rate the usefulness of the instructor's feedback on your performance.
6. Rate how well the graded materials (tests, papers, projects, etc.) reflected the course objectives.
7. Rate how well the objectives and requirements of the course were explained.
8. Rate the instructor's ability to make the course material understandable.
9. Rate how well the instructor stimulated interest in the subject.
10. Rate the overall quality of instruction in this class.

Of note is that 65% of the item responses were "Excellent." While most instructional experiences may in fact be "excellent," the well established "error of leniency" appears to be at work here, perhaps facilitated by the highest rating being listed first. Not evident from this table is that there is also a strong tendency for students to exhibit a response set bias or "halo effect" assigning the same rating (good or bad) to all ten items. In fact, for 42% of the SETE forms all ten questions received identical responses.

This led to the question of how many factors are represented by the ten items. A principal components factor analysis of both the raw ratings and the section item averages revealed that the SETE is indeed a single factor instrument with all items in the section analysis having a communality of at least .710 with that one factor. The item with the highest communality with this factor (.956) was item ten: "Rate the overall quality of instruction in this class." Thus, while the SETE provides a strong indicator of students' global evaluation, the current version appears to have limited value in assessing the underlying components of instructional effectiveness purportedly measured by these ten items.

Having a single factor instrument greatly simplified the subsequent analysis. A simple summative measure of teaching effectiveness was created for each section surveyed by determining the average of the ten item averages. This mean rating of the ten items for each section, while still exhibiting a slightly skewed distribution, served as the dependent measure for the analyses that follow.

Distribution of Mean Section Ratings

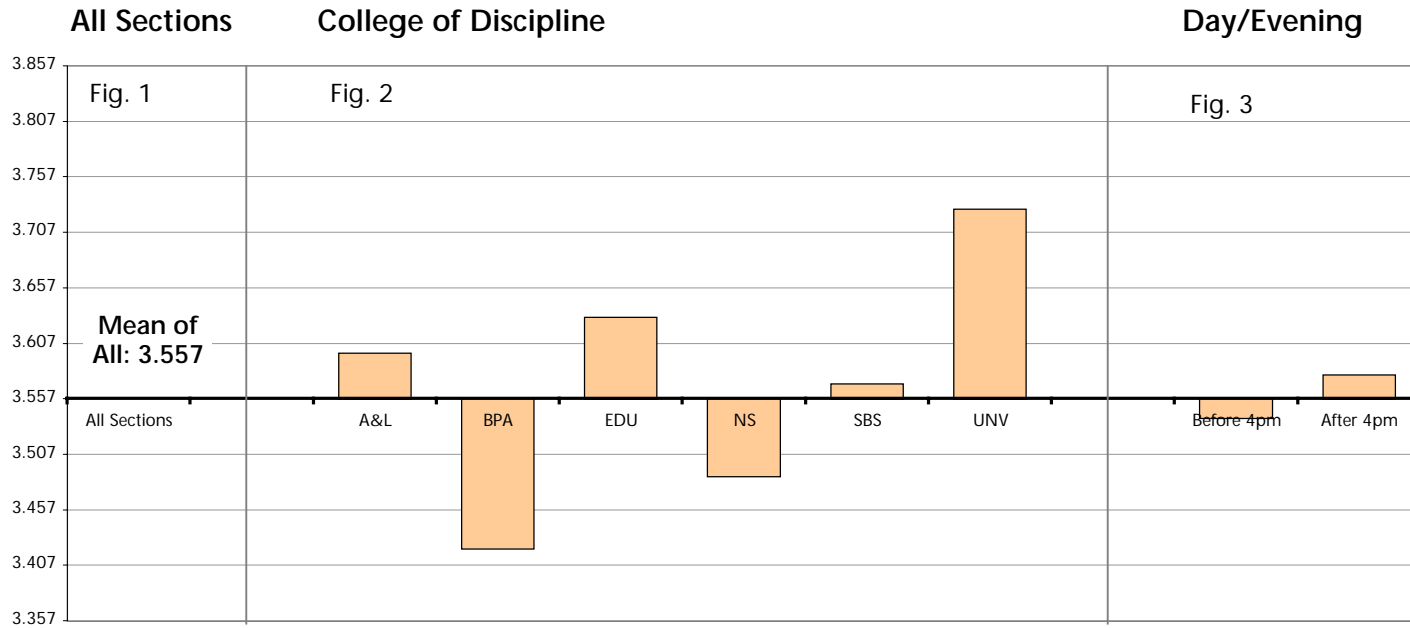


The initial impetus for this research was to determine what norms, in addition to College, might be used for SETE reporting. Through the use of APDB Section and Faculty files and grade distribution files, the following variables were examined for their role in course ratings:

- College,
- Class Level,
- Class Division,
- Start Hour,
- Day/Evening Class,
- Sessions per Week,
- Mode of Instruction,
- Instructor Rank,
- Class Size,
- Percent of Class Completing Form, and
- Average Letter Grade Assigned.

The latter three continuous variables were each clustered into an ordinal grouped variable for this stage of the analysis based on logical break points in the variable or the percentage of sections in a category.

Each of these independent variables was statistically significant in an analysis of variance. However, since these data were not derived from an experimental manipulation of instructional variables, the appropriate statistic for this analysis is the level of association between each of these variables and the summative course rating together with an assessment of the practical importance of magnitude of effect observed. The “Eta Squared” statistic indicates the percentage of variance in the summative course rating that is associated with the independent variable. These findings are presented in the pages that follow.



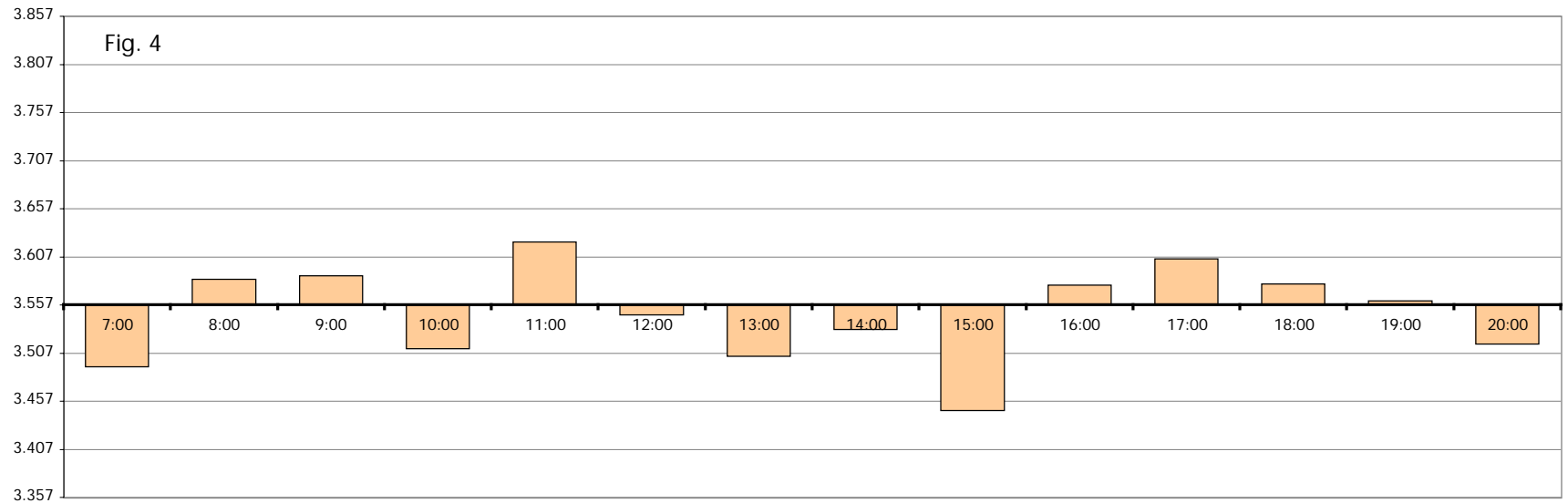
	All Sections	A&L	BPA	EDU	NS	SBS	UNV	Before 4pm	After 4pm
Mean	3.557	3.598	3.422	3.631	3.488	3.571	3.728	3.540	3.578
Sections	2,951	841	318	575	655	531	31	1,642	1,294
% of Sects.	100%	29%	11%	20%	22%	18%	1%	56%	44%
Std. Dev.	.341	.310	.357	.328	.362	.328	.346	.346	.335
Eta Squared		.042						.003	

Fig. 1: The chart and table given in Fig. 1, above, indicates the average rating across all sections of 3.557 or just above mid-way between “Good” and “Excellent.” The bold baseline for the graphs that follow all reference that average. The column charts indicate the average deviation from that mean for the sub-group analyzed.

Fig. 2: This replicates the current differentiation by College. Taken by itself, the College of Discipline was associated with 4.2% of the variability in SETE scores. Some of these observed differences may be due to other factors which differentiate Colleges.

Fig. 3: One alternative set of norms suggested would be to differentiate between day and evening classes. This distinction does not seem warranted given the 0.3% of variance associated with this dichotomy. A more detailed analysis by start hour is reported in Figure 4.

Start Hour-Military Time

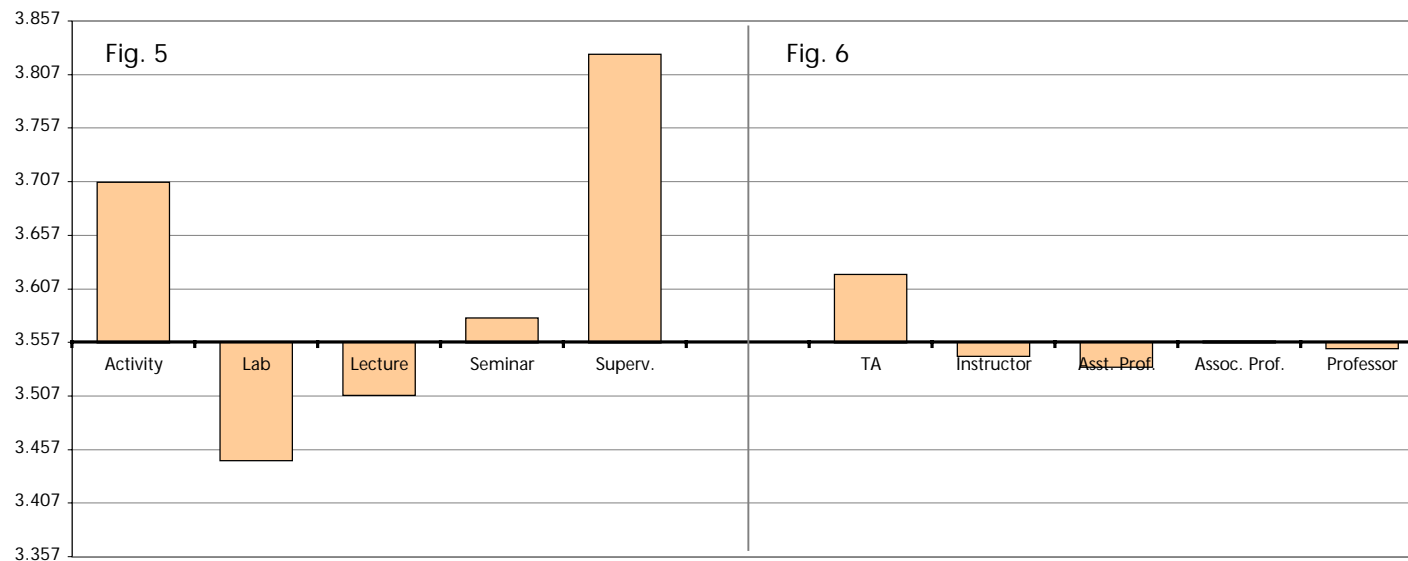


	7:00	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00
Mean	3.493	3.584	3.587	3.512	3.623	3.548	3.505	3.532	3.448	3.577	3.605	3.579	3.561	3.517
Sections	31	229	190	379	15	339	203	235	21	337	139	753	7	58
% of Sects.	1%	8%	7%	13%	1%	12%	7%	8%	1%	12%	5%	26%	0%	2%
Std. Dev.	.496	.344	.281	.325	.456	.357	.359	.353	.435	.306	.431	.330	.262	.315
Eta Squared	.009													

Fig. 4: There appear to be clear differentiations by the time of day that the class started. While this variable was associated with less than 1% of the variance, there appeared to be a disadvantage associated with classes that started in the 3 o'clock hour of the afternoon. (Note: This author observed the same mid-afternoon slump phenomena while revising a course evaluation instrument at the University of Oregon 25 years ago!)

Mode of Instruction

Instructor Rank



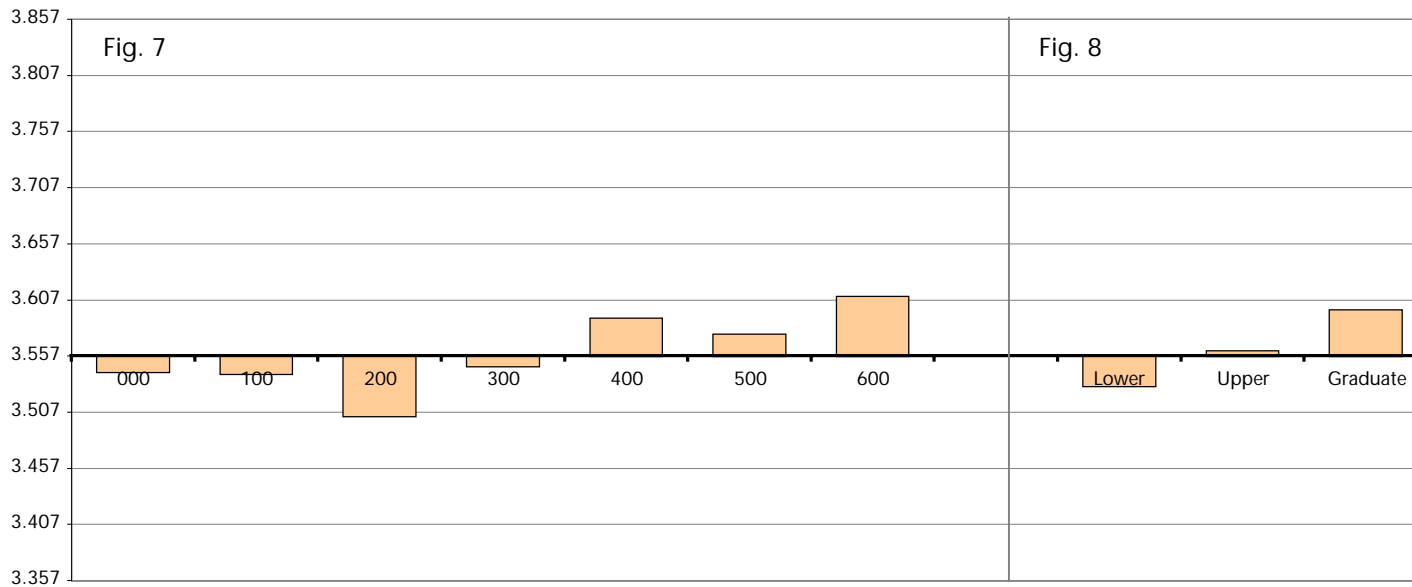
	Activity	Lab	Lecture	Seminar	Superv.	TA	Instructor	Asst. Prof.	Assoc. Prof.	Professor
Mean	3.706	3.447	3.508	3.580	3.826	3.620	3.545	3.535	3.559	3.552
Sections	169	124	1,087	1,543	28	489	1,315	515	203	429
% of Sects.	6%	4%	37%	52%	1%	17%	45%	18%	7%	15%
Std. Dev.	.274	.422	.336	.335	.337	.319	.351	.344	.333	.328
Eta Squared	.031					.007				

Fig. 5: Mode of Instruction was associated with 3.1% of the observed variance with laboratory experiences rating lowest, and the relatively rare supervision courses rating highest. Seminars rated higher than lectures, but that may have been a function of class size as discussed with Figure 9, below.

Fig. 6: Instructor Rank was associated with less than 1% of the variance. However, it is interesting to note that Teaching Assistants (TAs) are the most highly rated rank. One might have expected that ratings of the excellence of instruction would have increased with instructor rank. If this factor were to be used for norms, it would seem appropriate to group all tenure track faculty together and perhaps also include the Instructor (part-timers).

Class level

Division



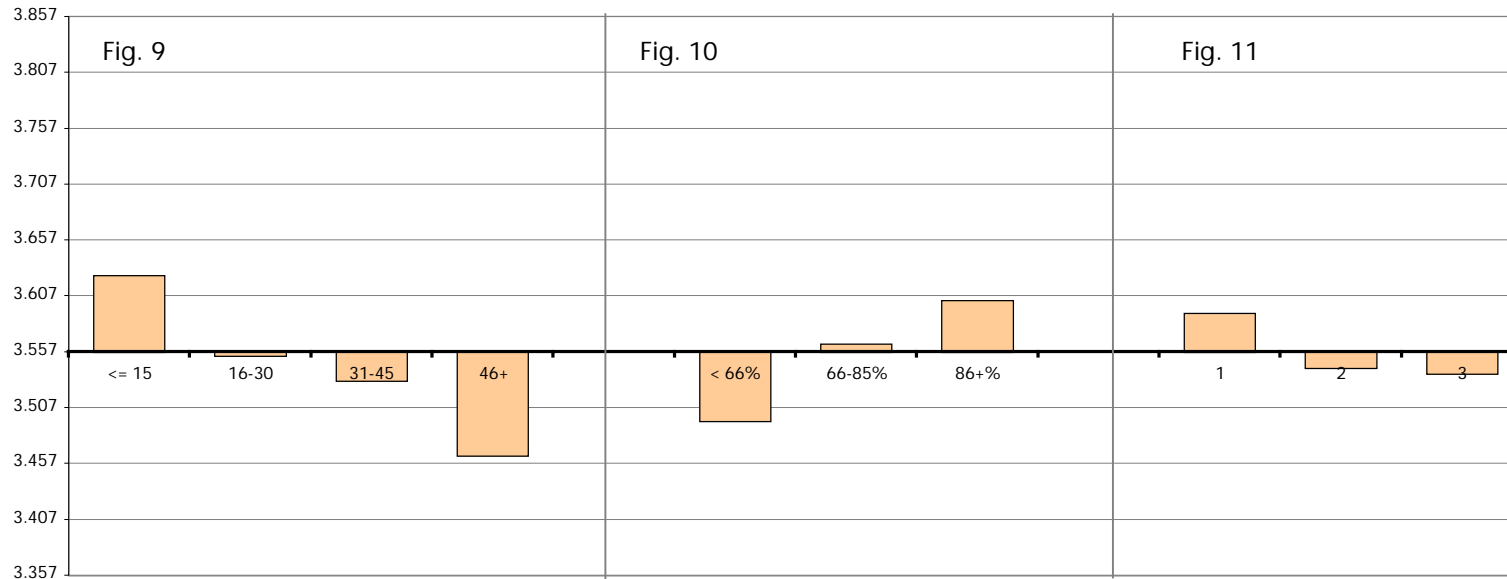
	000	100	200	300	400	500	600	Lower	Upper	Graduate
Mean	3.543	3.541	3.504	3.548	3.591	3.577	3.610	3.531	3.561	3.598
Sections	169	560	293	961	433	196	339	1,022	1,394	535
% of Sects.	6%	19%	10%	33%	15%	7%	12%	35%	47%	18%
Std. Dev.	.378	.318	.353	.329	.348	.382	.339	.339	.336	.355
Eta Squared	.008							.005		

Fig. 7 & 8: Both a higher course class level and grouping by division were associated with a higher SETE evaluation, though the variance associated was under 1%. While the comparison by division would seem to indicate a linear trend with higher level classes receiving higher ratings, the ratings of 200 and 500 level courses appear exceptions to this generalization.

Enrollment-Grouped

Pct. Completing SETE-Grouped

Sessions per Week



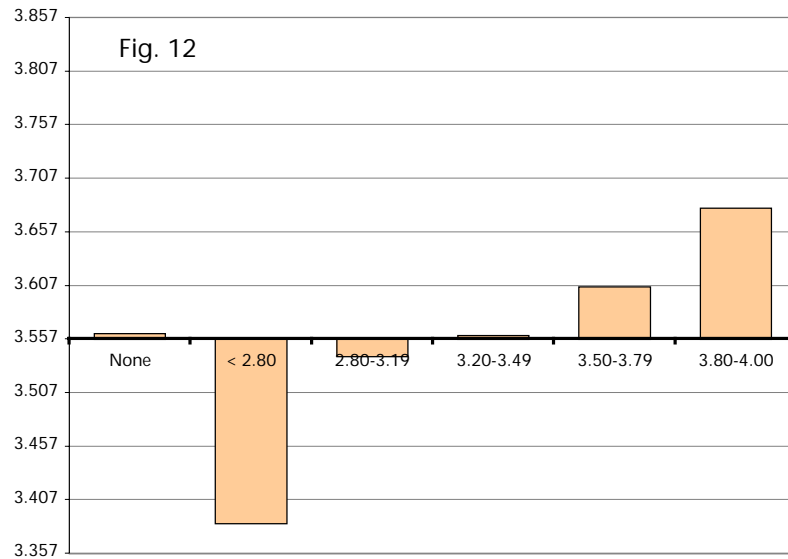
	<= 15	16-30	31-45	46+	< 66%	66-85%	86+%	1	2	3
Mean	3.625	3.553	3.531	3.464	3.495	3.564	3.603	3.592	3.543	3.538
Sections	709	1,495	414	333	735	1,408	808	933	1,440	566
% of Sects.	24%	51%	14%	11%	25%	48%	27%	32%	49%	19%
Std. Dev.	.355	.340	.306	.333	.358	.326	.343	.345	.341	.331
Eta Squared	.019				.013			.005		

Fig. 9: Another alternative report grouping was by class size. This analysis suggests that just under 2% of the variance in the total SETE rating was associated with class size. However, recall that Mode of Instruction accounted for half again as much variance.

Fig. 10: An interesting observation from a theoretical perspective was the 1.3% of variance associated with a grouping of sections by form completion rates. This measure was derived by determining the percentage of enrolled students who completed a SETE form. This served as an indicator of student attendance in class and possibly engagement in the instructional process.

Fig. 11: Classes which met once per week obtained higher ratings than classes which met two or three times weekly.

Letter Grade Average, Grouped



	None	< 2.80	2.80-3.19	3.20-3.49	3.50-3.79	3.80-4.00
Mean	3.562	3.385	3.541	3.560	3.605	3.679
Sections	236	489	607	519	573	527
% of Sects.	8%	17%	21%	18%	19%	18%
Std. Dev.	.365	.364	.321	.332	.315	.301
Eta Squared	.069					

Fig. 12: The single included variable with the strongest association with the average SETE rating was the average grade assigned in the course. Since the grades were not assigned until after the SETEs had been administered, direct causality is not possible. At least two explanations come to mind: Students may have had a “preview of coming attractions” based on their course grades to date so they would have had a reasonable estimate of the instructor’s grading standards and might thus assign lower ratings to more demanding instructors; or classes where the students were engaged in learning were characterized by both higher grades for students and higher evaluations for instructors.

Multiple Regression

A more sophisticated approach to estimating the impact of these variables upon the summative course rating is multiple regression.

The first analysis considered all of the input variables that are characteristics of a course: College of Discipline, Class Division, Mode of Instruction, Instructor Rank (recoded to TA, Part-Timer and Tenure Track), Class Size (grouped), Sessions per Week, and Day/Evening Classes were initially analyzed.

Considering those variables which were both:

- statistically significant at the .05 level, and
- with a practical magnitude of effect of at least a 0.1 difference in average rating between the highest and lowest values of a variable
- excluded all but Mode of Instruction and Class Size. The summary of the regression is attached.

Variations due to class size can be attributed to the mode of instruction since standards for assigning a class to an instructional mode do factor in the number of enrollments allowed. If class size parameters for mode of instruction are followed, then Mode would appear to be a useful basis for norm grouping which would implicitly consider class size. However, since many seminar class are sized as lectures, Class Size is still a useful variable. Further, with the possible exception of the College of Business and Public Administration, there does not appear to be a justification for creating separate norms for each College.

Grade Average and Percentage Completing a SETE were not considered because neither were course attributes. As noted above, both were nevertheless associated with SETE ratings and may of interest in a follow-up analysis

Minor Recommendations

Modify Alternatives: One solution to the skewness problem whereby ratings bunch at the high end would be to merge the little used two lowest ratings which together account for less than 10% of the responses, rewording the descriptor to "Fair or Poor," and adding a "Very Good" rating between "Excellent" and "Good."

Norms: The above analyses show some support for creating additional norm groupings based on Class Size and Mode of Instruction. Optionally, college specific norms could be dropped.

Rethinking the System

The current single form SETE could be abandoned in favor of two different course evaluation instruments with differing objectives.

Formative: The first would be used to facilitate formative evaluations of teaching effectiveness. The items would document the extent to which students observed the instructor engaged in specific agreed-upon desirable instructional behaviors. Different sections of the form would address mode specific expectations. (E.g., unique to lab activities.) These ratings would not be inherently comparative or qualitative and enrollees would serve as witnesses, not judges.

Summative: The second would provide a comparative rating. Although unwilling or unable to discriminate among levels of instructional performance, students do appear able to provide overall evaluations that are unfortunately currently limited by the clustering of responses at the positive end of the response scale. The simplest way to address this issue is to expand the high end of the scale as indicated above. A more sophisticated alternative would be to implement a forced-choice approach where students would rank their courses from high to low based on teaching effectiveness. Since each student would require a customized list of courses taken, this method is not currently practical. However, with the advent of web-based registration, comparative course evaluations could be obtained for the student's current courses when registering for the subsequent term. Each registering student would be given a list of three courses randomly drawn from those in which they are currently enrolled. (Listing a course or two from previous terms if they are enrolled for less than three.) Students would identify the one most effectively taught course of the three. This method would:

- expand the base of courses/instructors evaluated,
- give each student an equal "vote,"
- facilitate comparisons within programs since courses taken (and rated) by students tend to cluster within programs, and
- effectively identify the top third of instructors, while
- avoiding the labeling of any as the "least effective."

The section data used for this report is available in SPSS format for interested and qualified members of the campus community who may wish to further research these issues.