GRINNELL COLLEGE

Institutional Research

Project Report
October 1999

Carol Trosset, Director
Scott Baumler, Senior Research Analyst

Fell House, Grinnell College
Grinnell, Iowa 50112
515-269-4931

End-of-Course Evaluations: Spring 1999 Results

Executive Summary

Statistical analyses of the end-of-course evaluation form used in the spring of 1999 suggest that the instrument was reasonably reliable and unbiased. However, qualitative analysis of text comments showed that students used varying criteria many of which did not relate directly to the questions asked about their own learning. This suggests that the form is better suited to measure satisfaction than student learning or teaching quality.

Any attempt to calculate or compare summary scores should utilize statistical margins of error (confidence intervals) to avoid overemphasizing small numerical differences. This severely limits the number of distinctions that can be drawn between different courses or professors, resulting in broad categories like very satisfied, satisfied, and not very satisfied.

If a measure of satisfaction is what is desired this form is probably adequate, though the wording of the questions could be improved to fit better with how students tend to phrase things. It is important, however, not to assume that what the students are satisfied with is always the quality of teaching. Qualitative data from comments have given us a list of factors deemed important by many students. Only if all these factors were also considered important and valid measures by most faculty members, or if student scores were found to correlate significantly with independent measures such as peer review or learning outcomes, could we claim to have a valid measure of teaching quality.

Overview

For several years prior to spring semester of 1999, end-of-course evaluations were mandatory but each department used its own form. Many departments had more than one form for use with different courses. During 1998, Laura Sinnett analyzed responses from faculty members who agreed to share their evaluation data from the spring of 1998. Sinnett identified questions that were asked in a similar fashion across several departments, and performed statistical analyses to test the reliability and validity of the instruments. On the basis of Sinnett's results, and taking into consideration the large amount of human and other resources going into the end-of-course evaluation process, Executive Council in the spring of 1999 recommended a one-semester test of a single college-wide form.

The questionnaire had six evaluative questions which were developed by the Executive Council and adopted at the March 15, 1999 faculty meeting:

Q1: Activities during our meeting time significantly contributed to my learning.

Q2: Interactions with the instructor (either inside or outside of the meeting time) significantly contributed to my learning.

Q3: Interactions with the other students (either inside or outside of the meeting time) significantly contributed to my learning.

Q4: Oral and written work, tests, and other assignments significantly contributed to my learning.

Q5: Course materials (for example, readings, films, and studio, laboratory, or activity manuals and equipment) significantly contributed to my learning.

Q6: Overall, this course significantly contributed to my learning.

A machine-scannable form was constructed to gather the information. The six-point scale employed was anchored by Strongly Disagree, Moderately Disagree, Slightly Disagree, Slightly Agree, Moderately Agree, and Strongly Agree. Space was also provided on the form for text comments.

Five additional questions were included to gather data about each student's year in college, gender, anticipated grade in the course, likely GPA effect, and how the course related to the student's major.

Data were gathered from 306 course sections. Overall, students tended to use the upper end of the rating scale. Across all six questions, 45 percent of the valid responses were Strongly Agree. Subsequent tables provide summary statistics, where the responses were coded as follows:

Strongly Disagree	Moderately Disagree	Slightly Disagree	Slightly Agree	Moderately Agree	Strongly Agree
1	2	3	4	5	6

Table 1: Summary of Student Responses

	Q1	Q2	Q3	Q4	Q5	Q6
Average Response	5.19	5.24	4.79	5.04	5.07	5.28
Std. Deviation	1.05	1.05	1.19	1.09	1.11	1.03

Table 1: Summary of Student Responses (continued)

Frequencies

Q1: Activities during our meeting time

	Frequency	Percent
Strongly disagree	68	1.5
Moderately disagree	106	2.3
Slightly disagree	124	2.7
Slightly agree	489	10.6
Moderately agree	1,595	34.4
Strongly agree	2,222	47.9
Not applicable/don't know	24	0.5
No response	7	0.2
	4,635	100.0

Q2: Interactions with the instructor

	Frequency	Percent
Strongly disagree	70	1.5
Moderately disagree	78	1.7
Slightly disagree	129	2.8
Slightly agree	483	10.4
Moderately agree	1,347	29.1
Strongly agree	2,293	49.5
Not applicable/don't know	231	5.0
No response	4	0.1
	4,635	100.0

Q3: Interactions with other students

	Frequency	Percent
Strongly disagree	95	2.0
Moderately disagree	172	3.7
Slightly disagree	211	4.6
Slightly disagree	985	21.3
Moderately agree	1,449	31.3
Strongly agree	1,395	30.1
Not applicable/don't know	321	6.9
No response	7	0.2
	4,635	100.0

Q4: Work, tests, assignments

	Frequency	Percent
Strongly disagree	70	1.5
Moderately disagree	105	2.3
Slightly disagree	177	3.8
Slightly agree	666	14.4
Moderately agree	1,621	35.0
Strongly agree	1,777	38.3
Not applicable/don't know	212	4.6
No response	7	0.2
	4,635	100.0

Q5: Course materials

	Frequency	Percent
Strongly disagree	81	1.7
Moderately disagree	114	2.5
Slightly disagree	158	3.4
Slightly agree	629	13.6
Moderately agree	1,535	33.1
Strongly agree	1,899	41.0
Not applicable/don't know	195	4.2
No response	24	0.5
	4,635	100.0

Q6: Overall

	Frequency	Percent
Strongly disagree	71	1.5
Moderately disagree	80	1.7
Slightly disagree	108	2.3
Slightly agree	460	9.9
Moderately agree	1,427	30.8
Strongly agree	2,470	53.3
Not applicable/don't know	9	0.2
No response	10	0.2
	4,635	100.0

Course-Level Results

Below are summary statistics for mean scores at the course level (i,e., class averages as the unit of analysis). Please note that when the term "class" or "course" is used in this report it refers to an individual course-section.

Table 2: Course-Level Summary Statistics

Statistic	Q1	Q2	Q3	Q4	Q5	Q6
Average Score	5.26	5.28	4.85	5.12	5.12	5.34
Standard Deviation	0.49	0.51	0.55	0.47	0.58	0.47
Range	2.18	3.50	3.25	2.37	3.30	2.50
25th Percentile	4.96	5.00	4.50	4.82	4.82	5.08
Median	5.32	5.38	4.89	5.17	5.20	5.40
75th Percentile	5.65	5.64	5.23	5.45	5.50	5.67
Number of Courses	306	306	305	304	306	306

Reliability

Reliability generally refers to measurement consistency. Methods frequently used to assess reliability employ test-retest, alternative form, and parallel-measure approaches. While the design of last spring's experiment did not allow for these particular assessment methods, the methods that were available did indicate stability in the measures.

Confidence Intervals

Confidence intervals denote margins of error. Providing intervals – ranges around mean scores where the "true" values are likely located – is important for communicating an accurate sense of the quality of the results. The use of intervals actively recognizes chance error in survey results and guards against misplaced confidence in singular point values.

The ranges were generally wider for courses with small enrollments, which would be expected given that confidence intervals depend not only on the variability of the data but on the number of observations as well. Statistically, little could be said about the mean scores for many courses with fewer than ten students. For instance, in three courses with small enrollments, it could be said with a high degree of certainty that the mean student rating for the course overall was somewhere between strongly disagree and strongly agree (i.e., the confidence interval spanned the entire scale).

Differentiation

A reliable instrument can identify reliable differences. To interpret differences in score profiles, one must be able to discern whether the differences could have resulted from mere chance. Confidence intervals can be used to compare the number of likely "real" distinctions among course

mean scores. As in Sinnett's material, confidence intervals were compared across courses to count the frequency of non-overlapping intervals. When the intervals do not overlap the scores probably reflect genuine differences. When the intervals overlap considerably no importance should be attached to differences between the scores. The procedures were conducted in the same manner as the previous research to facilitate comparison. Please note these are 95% confidence intervals for each mean *separately*.

The procedure was carried out by first calculating the confidence interval for Q1 for each course. The interval for Course #1 was then compared to the interval for Course #2 to check if the intervals overlapped. Course #1 was then compared to Course #3, Course #4, Course #5, and so on until all possible pairwise comparisons were made.

The second round of the routine compared Course #2 to Course #3, Course #4, Course #5, and so on. This continued for 306 rounds (as there were 306 course sections), resulting in 46,665 combinations (adjustments were made for comparisons already carried out). The remaining five questions were then analyzed in the same manner. Table 3 provides the results.

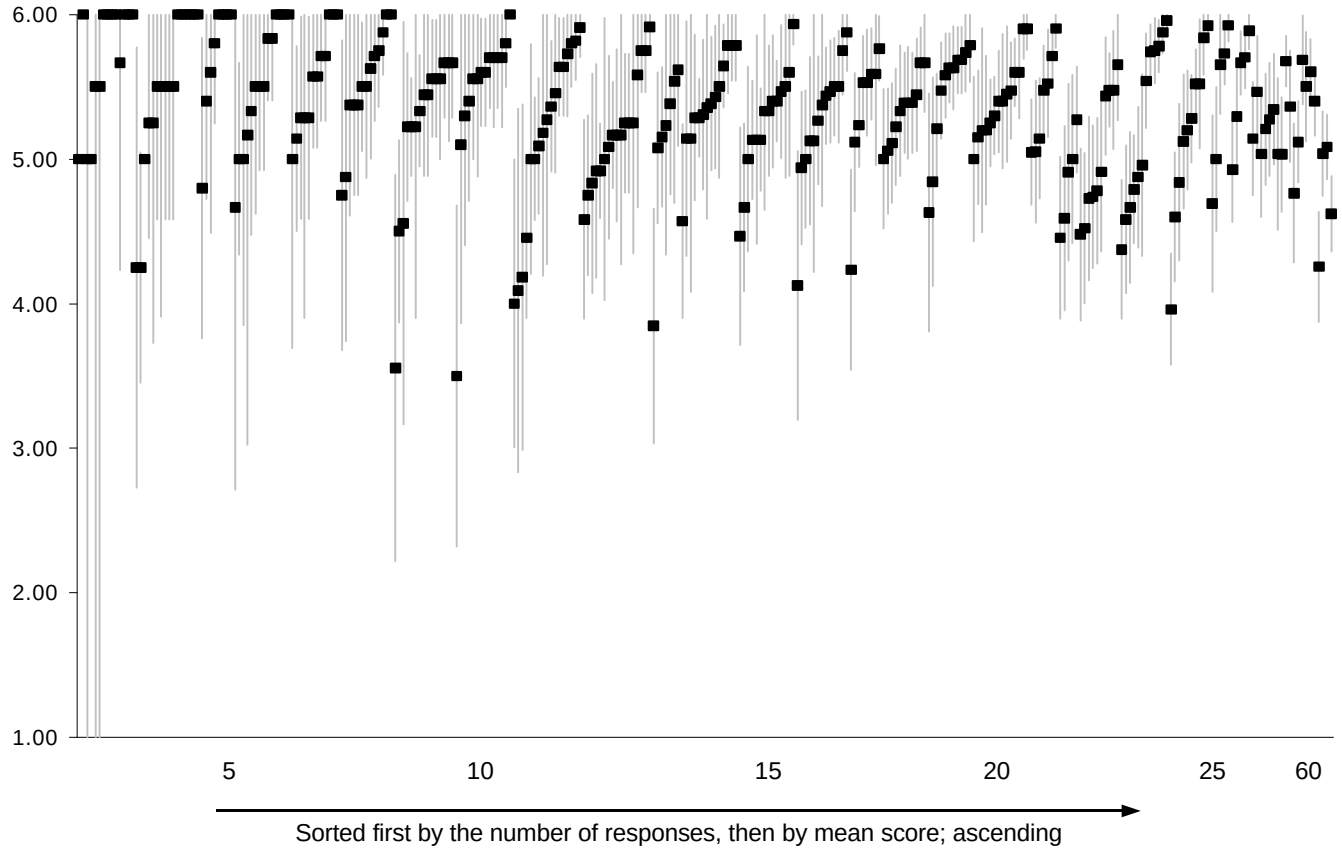
Table 3: Pairwise Comparisons of Confidence Intervals by Course

Question	Confidence Intervals Did Not Overlap	Confidence Intervals Did Overlap	Total Number of Comparisons	Percent That Did Not Overlap	N
Q1	9,203	37,462	46,665	20%	306
Q2	8,271	38,394	46,665	18%	306
Q3	6,204	40,156	46,360	13%	305
Q4	6,422	39,634	46,056	14%	304
Q5	8,619	38,046	46,665	18%	306
Q6	9,815	36,850	46,665	21%	306
Total	48,534	230,542	279,076	17%	

The differentiating power of the new form was on par with the sample of end-of-course forms analyzed in Phase I. In both sets of data, 17 percent of the comparisons overall resulted in non-overlapping confidence intervals. Absent an objective criterion against which to gauge the adequacy of this result, it is difficult to know whether 17 percent is good or poor. It at least indicates stability between the two different end-of-course evaluation approaches (the prior distributed system versus the uniform questionnaire under review).

Figure 1 illustrates Q6 mean scores and confidence intervals for all 306 course sections. The x-axis is categorical; each item on the horizontal axis represents a course. Numbers below the x-axis indicate class size. Mean score values are on the y-axis. The small black boxes are mean scores and the vertical lines represent the confidence intervals. Patterns in the graph appear because the courses were sorted first by the number of survey responses then by mean score. The average range (numerical span) of the Q6 confidence interval was .95 points. Over all six questions, the range averaged 1.12 points.

Figure 1: Mean Scores & Confidence Intervals for All Courses, Question #6



The interval-comparison procedure outlined above was also used to analyze the data when it was grouped by instructor rather than by course. Pooling the data by person provides a more direct view of the differentiating ability of the instrument for summative purposes.* The appropriateness of combining scores across courses and the usefulness of all six questions for drawing general distinctions could be debated. Be that as it may, the material is presented in Table 4 for exploratory purposes.

The procedure was also carried out by instructor with the confidence intervals set at 90 percent rather than 95 percent. The percent of intervals that did not overlap ranged from 23 percent for Q4 to 34 percent for Q1. Thirty-one percent of the comparisons for Q6 did not overlap and the overall rate (for all six questions) was 29 percent.

* Class average scores could also be aggregated by instructor for this type of analysis. This would provide an equal weighting of courses versus and equal weighting of responses. Due to the small number of cases available for each instructor at the course level this test was not conducted.

Table 4: Pairwise Comparisons of Confidence Intervals by Instructor

Question	Confidence Intervals Did Not Overlap	Confidence Intervals Did Overlap	Total Number of Comparisons	Percent That Did Not Overlap	N
Q1	2,971	8,204	11,175	27%	150
Q2	2,425	8,750	11,175	22%	150
Q3	1,817	9,358	11,175	16%	150
Q4	1,716	9,310	11,026	16%	149
Q5	2,721	8,454	11,175	24%	150
Q6	2,603	8,572	11,175	23%	150
Total	14,253	52,648	66,901	21%	

Consistency Measures

A reliable instrument will provide consistent results from repeated measures of the same phenomenon. Precise replication is unlikely due to random error, though a good device will minimize chance error to yield answers that reflect only true variation. The amount of agreement among measures can be used to gauge reliability.

For example, the consistency with which different judges score the same performance speaks to reliability. Good performances should consistently receive high scores and weak performances should consistently receive low scores. In this context, reliability can be measured in terms of inter-rater agreement.* The intraclass correlation coefficient (ICC) is a statistical indicator that can be used for this purpose.

An ICC is a measure of homogeneity. In an analysis-of-variance fashion, an ICC compares different sources of variation. The indicator approaches 1.0 when the variation between subjects is large relative to the variation within subjects. That is, an ICC close to 1.0 indicates a very consistent measure.

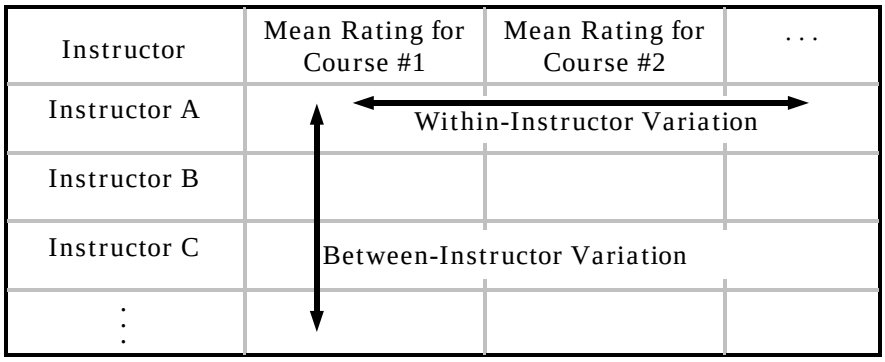
The basic data layout is depicted in Figure 2. In panel a of the figure, instructors are listed down the left-hand side in the rows of the grid. The different courses taught by the instructors are in the columns. The numbers placed inside the respective grid cells are mean scores for a particular survey question. If there exist certain teaching effectiveness qualities that an instructor possesses (aspects the students rate highly), one would expect to find relatively consistent scores across the rows. Specifically, if there is more consistency across the rows than down the columns this is considered an instructor effect (the order the courses are listed in the columns becomes irrelevant with this particular application).

Panel b of the figure represents course effects. The rows in this grid contain parallel courses (same subject and course number; different section) taught by different instructors. If the course itself is what drives the mean score, consistency in the rows should be evident. Panel c represents the few instances where the same instructor taught the same course twice (different

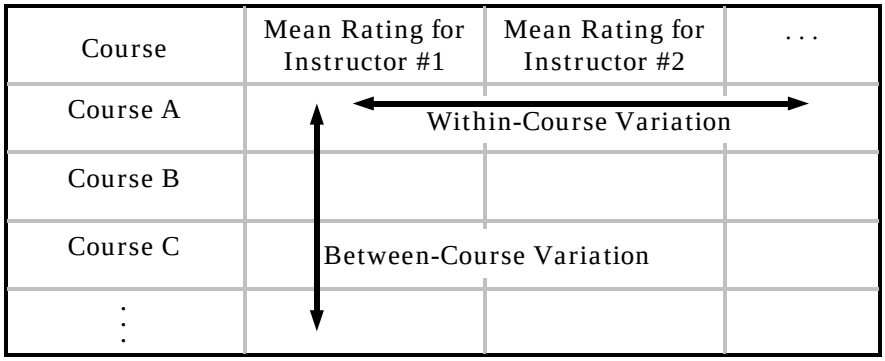
* *SPSS Base 9.0 Applications Guide*, 1999, SPSS Inc.

Figure 2: Data Structure for Consistency Measures

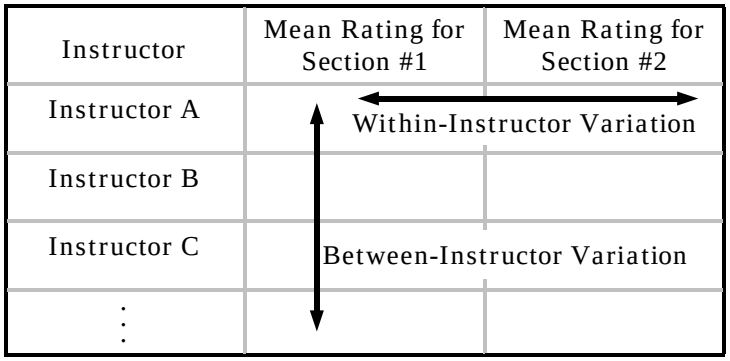
a) Same instructor who taught multiple different courses



b) Same course taught by different instructors



c) Same instructor who taught two sections of the same course



sections) in the spring of 1999. This was the closest to a quasi-experimental design that could be found in the data for this purpose.

One caveat for this approach concerns the sample composition of these groups. Panel a, the instructor effect group, mainly includes people who taught an upper-level and a lower-level course. Panel b, the course effect group, contains mainly introductory-level courses. Three of the eight cases in the "same course, same instructor" group are laboratories. The composition is simply an artifact of the available data.

When differences owing to row comparisons (between-group variation) are large with respect to differences derived from column comparisons (within-subject variation), the ICC gets larger. The ICC will tend toward its maximum value when the rows have the same score in each column. The results for this set of information are not so pronounced, though the statistics do provide some intriguing evidence. *Instructor effects do appear to outweigh course effects. The actual differences among the means are, however, relatively small.* Table 5 provides the data.

Table 5: Consistency Statistics

Intraclass Correlation Coefficients

	Same Course, Different Instructors		Same Instructor, Different Courses		Same Course, Same Instructor	
Question	**Average Measure ICC**	**# Cases**	**Average Measure ICC**	**# Cases**	**Average Measure ICC**	**# Cases**
Q1	0.45	30	0.52	104	0.77	8
Q2	0.20	30	0.60	104	0.92	8
Q3	0.35	30	0.39	103	0.85	8
Q4	0.27	30	0.52	103	0.60	8
Q5	0.62	30	0.58	104	0.92	8
Q6	0.34	30	0.52	104	0.76	8

Mean Absolute Deviations

	Same Course, Different Instructors		Same Instructor, Different Courses		Same Course, Same Instructor	
Question	**MAD**	**MAD %**	**MAD**	**MAD %**	**MAD**	**MAD %**
Q1	.25	4.8	.23	4.5	.13	2.6
Q2	.23	4.5	.23	4.5	.06	1.2
Q3	.28	6.1	.28	5.9	.12	2.4
Q4	.25	5.0	.22	4.3	.17	3.3
Q5	.25	5.4	.27	5.5	.08	1.6
Q6	.24	4.7	.21	4.0	.12	2.2

The bottom portion of Table 5 lists mean absolute deviations. These statistics indicate the "tightness" of the measures; the deviation of each element from its own mean. The numbers were derived by calculating the mean for each target (row) item and subtracting it from each element in the row. The absolute value of these differences were then averaged.

There are multiple methods for this type of assessment, but this procedure preserved the actual point value to demonstrate the variability. As a general example, take a case where one observed measurement equaled 4.0 and another equaled 6.0. The average of these two values is 5.0, so the mean absolute deviation would be one point.

Turning to the data in the Table 5 for an example, note the mean absolute deviation of 0.13 for Q1 in the right-hand column. The cases where the same instructor taught the same course yielded about a quarter-point spread between measurements on this item (each case in this group had only two observations).

Validity

The issue of validity addresses whether a question measures what it was designed to measure. There are many aspects to validity, including generalizability, biases, corroboration with other measures, predictive ability, the domain of factors affecting the measures, and the uses to which the results are put. Not all of these aspects were able to be fully addressed given the parameters of this study, but the following information touches on many of the themes.

Confounding Variables

Given the power available with a sizable sample, a number of confounding factors could be deemed statistically significant. The effect sizes of these factors, though, were generally small. With the continued use of confidence intervals, subtle point differences may not be particularly troublesome for general summative purposes. *Statistical* significance may differ from *practical* significance in this sense.

Tests and breakdowns can be applied ad nauseam to this data set. Tables in the appendix were developed in a general sweep to address basic issues and to build on the validity analysis material presented in the Phase I pilot study. In general, upper-level courses and courses with smaller enrollments received higher marks. Female students typically gave higher scores than males. Students who expected to receive lower grades gave lower ratings. The same is true for students who expected a negative impact on their GPAs. Students for whom the course was in their actual or intended field of study gave higher scores on average than did students for whom the course was not in their field of study or had not yet declared a major.

Validity Concerns

Four percent of the sample investigated misinterpreted the wording of the first three questions. "Activities during the class meeting" was taken by these students to mean only things other than lecture or discussion. "Interactions with the instructor" was taken to mean only face-to-face interaction outside of class (even though "in or out of class" was specified in the text of the question). Students making these errors tended to disagree with the statements rather than to mark "not applicable," thereby lowering the professor's average scores.

What students really think about when they evaluate a course or professor

Unlike interviews, where the interviewer can ask follow-up questions to keep the subject focused on the intended topic, surveys have no way of forcing the respondent to answer the question that was posed. Comments provide our only indication of what students really think about when they fill out course evaluations. Obviously, not all their thoughts are described in comments, but we do know that anything they commented on is something they thought about. By collecting many comments written by many students, and grouping them by topic, we can develop a list of the factors students (as a population, not necessarily as individuals) think about when evaluating courses and professors.

Identifying these factors enables us to evaluate the "content validity" of the questions asked. All the questions on the Spring 1999 form asked the students to focus on their own learning, but most of the student comments focused on other things. Of a sample including 1,933 distinct comments that had some direct reference to the professor:

- 32% were about personal attributes (such as niceness, energy, or availability)
- 30% were about helpfulness (as in "helped me understand the material better"--indirectly about learning)
- 26% were about perceived competence (such as knowledge level or whether they liked how the class was run)
- 12% of the comments made a direct reference to whether the student had learned anything.

This kind of content analysis can be used to build a more detailed profile of what things are on students' minds when they evaluate courses and professors. The items on the list below were mentioned (positively or negatively) again and again on the Spring 1999 forms. Given that we surveyed about a thousand students four times each, and the content analysis was based on a 40% sample including courses in each department at each level, this list can be thought of as a composite student model of what constitutes good teaching. Not every student will be thinking about all of these factors, but a typical group of students, such as might turn up in most courses, would probably consider this range of things when they complete course evaluations. As we saw with the Spring 1999 responses, students are likely to consider these factors regardless of exactly what questions are asked on the forms.

- ✓ professor availability
- ✓ professor niceness or approachability
- ✓ professor energy or enthusiasm
- ✓ apparent professor knowledge level
- ✓ how well the class sessions were run
- ✓ whether the student likes the chosen classroom format
- ✓ whether the professor helped the student understand the course materials better
- ✓ whether the course made the student think
- ✓ whether the student's skills increased

Imperfect correlations between numbers and comments

Many studies have been done at other institutions about course evaluations. Some studies have tried to correlate student ratings on evaluations with independent measures of teaching effectiveness like student exam scores, peer or administrator ratings, or instructor self-ratings. Some of these studies have found a significant positive correlation between these factors, suggesting that student ratings might actually be a valid measure of teaching quality. Others have found that the items which students rate the most reliably have little to do with the quality of instruction. Therefore, we should be cautious about using student ratings to measure teaching effectiveness at Grinnell.

Despite the fact that students do not write down all their thoughts, the range of topics reflected in the comments accompanying a particular score shows that the same score is not always awarded for the same reason. (The comments listed below are representative of the range things students said, but the order does not imply anything about the frequency with which different topics were mentioned.)

6 = STRONGLY AGREE

- My writing has improved after working on papers for this class.
- He did an incredible job of explaining complex information to students who haven't much background. Evidence of this fact: people in class consistently asked good questions.
- She always makes time--even on weekends. She gives us her life. She is wonderful.
- I always enjoyed talking to him outside of class — he took a real interest in my life.
- She is an intellectual powerhouse. All uncertainties and ambiguities were eliminated by her from the readings.

5 = MODERATELY AGREE

- She was well-organized during class and brought up a lot of important points.
- Discussion during the class period was a time during which the instructor could share knowledge of the text, which I could not obtain. Comments on outlines for papers were useful, good feedback.
- Yes, regular weekly meetings helped a lot. I felt more comfortable with the coursework.
- He was quite good at answering questions and easy to get in touch with outside of class. My only comment would be that every once in a while he seemed a bit hostile when asked a stupid question.
- He did a wonderful job of shaping the class discussion around the concerns of the student.
- I thought she could have been a bit more receptive to opposing viewpoints sometimes.
- He liked to meet with us about presentations, and I met with him about my paper. It just usually took at least 45 minutes and I thought it could have been 15. But he got me thinking I guess.
- Much of the learning came from reading the texts but he did wonderfully help us examine those texts and authors.

4 = SOMEWHAT AGREE

- I don't really feel like my prof really was what I learned stuff from. I mean, except for facts about the subject. (The readings taught me <u>tons</u> about lots of things I know nothing about.)
- I do feel my knowledge increased substantially, but I would have liked a deeper analysis of certain topics. Outside meetings were helpful though.
- She is obviously knowledgeable, and was much more approachable and helpful one on one than in class.
- Not very helpful, doesn't explain well.
- I feel like there was a lot more potential to this course than we ever really explored. We read case studies. I want theory.
- I feel like my ideas were not all the time validly listened to. Many times it seemed that he wanted you to do this <u>his</u> way, which is frustrating when you have your own ideas.

3 = SOMEWHAT DISAGREE

- 'Significantly' is a strong word here. I don't feel that I would have lost a great educational experience if I would have had a different instructor.
- Because much of class time was taken up with lecture, my attention was not always held. I frequently wished more class time was used for group discussion and focus on what we read for outside homework.
- I did not feel any encouragement from her. I did not feel welcome in office hours, and I felt she was not helpful with personal matters.

- Very nice, but not extremely helpful out of class.

2 = MODERATELY DISAGREE

- I was frustrated that the focus of this course centered on the dissection of details, rather than the discussion of events and narrative that occurred in the text we read.
- Interactions with my professor were mostly useless, mostly because of scheduling conflicts. I never left an interaction understanding material, nor feeling good about the situation.
- While the activities such as worksheets illustrated some basic features of the lesson, they were oversimplified, and the instructor always left the students to figure out the main part of the concept independently.

1 = STRONGLY DISAGREE

- I feel like when I asked for help I was simply told what I was supposed to be doing in broad terms, but I didn't understand how to do it.

Student perceptions of different types of courses

A cluster analysis enabled us to identify groups of courses with similar patterns of numeric responses to all six questions. Four clusters were readily identified using only the numeric responses. The "reality" of these clusters was verified when it turned out that courses in each cluster had a distinctive pattern of student comments. This suggests that there are four "types" of courses as perceived by students. About half of all faculty members consistently appeared in a single cluster; the other half had their courses scattered across (usually two) clusters.

Seventy-five percent of the courses fall into the two biggest clusters, which might be thought of as corresponding to stronger and weaker courses. They both contain quite representative distributions of classes (across divisions, levels, etc.).

a) The Typical Good Class. This includes 42% of the courses and receives the second highest mean overall score (5.6). The class is well run, the professor is helpful, and the other students and the course materials are good.

b) The Ambivalent Student. 33% of the courses fall into this cluster, which receives the third highest mean score (5.1). The distinctive feature of this cluster is the fact that almost every student expresses ambivalence about the course and/or the instructor. Some readings were good and others were not; the professor has this good quality and that bad one; some activities worked and others didn't.

The other 25% of the courses fall into the other two clusters, which received the highest and lowest mean overall scores.

c) The Charismatic Professor. This group includes 12% of the courses and receives the highest mean rating for the course overall (5.9). Comments consistently indicate that the professor is seen to have special personal attributes (extra nice, extra available, extra energetic). The class is well run, the professor is knowledgeable, and is given credit for assigning good reading materials. The other students are good, and they bond together and discuss the material outside of class. The course is seen as meaningful/relevant to the students' lives.

d) The Controversial Class. This cluster includes 13% of the courses and gets the lowest overall mean scores (4.5). This low mean, however, results from a bimodal distribution of scores. In these courses, some students think the course and professor were excellent while others think they were terrible. The latter group are numerous enough to pull down the average score, and often dislike how the class sessions were run. Criticism of other students is most common in this cluster.

These two clusters are more distinctive than the two larger groups. Courses appearing in the Charismatic cluster were disproportionately from the upper levels (200s and 300s), and their professors and students were disproportionately female. Nearly all courses in this cluster come from the Humanities and Social Studies divisions. Courses from the Controversial cluster tend to be more from the lower levels (mainly 100s, some 200s).

Further analysis would be needed to account for membership in these clusters, but anecdotal evidence suggests that one factor is probably related to student expectations about how a particular subject will be approached. The existence of these two clusters emphasizes the need to be cautious about assuming that high and low scores necessarily correspond to good and bad teaching.

Appendix

Please note the following conventions:

N refers to the number of observations.

F refers to an F-test

t refers to a t-test

R^2 and Eta^2 can both be interpreted as the proportion of total variation accounted for by the independent variable. R^2 assumes linearity, Eta^2 does not. (R^2 is listed when the categories represent an order or progression; the variables were coded accordingly.)

Sig. refers to the level of significance. Standard practice generally recognizes values of .05 or less as acceptable significance levels. Values listed as "0.00" are not actually zero but are small enough so as not to round up to 0.01.

Post-hoc (multiple comparisons) tests were completed using the Tukey HSD. Superscripted letters refer to the line item for which a comparison indicated a significant mean difference at the 0.05 level.

Pearson correlation coefficients are marked with asterisks to indicate significance levels. One asterisk (*) indicates significance at the 0.05 level (2-tailed) and two asterisks (**) indicates significance at the 0.01 level (2-tailed).

Table A-1: Pearson Correlation Coefficients for the Student Responses

	Q1	Q2	Q3	Q4	Q5	Q6
Q1	1.00 n = 4,604					
Q2	0.63** n = 4,372	1.00 n = 4,400				
Q3	0.39** n = 4,280	0.36** n = 4,142	1.00 n = 4,307			
Q4	0.48** n = 4,389	0.48** n = 4,200	0.32** n = 4,123	1.00 n = 4,416		
Q5	0.46** n = 4,389	0.40** n = 4,201	0.30** n = 4,117	0.41** n = 4,267	1.00 n = 4,416	
Q6	0.70** n = 4,586	0.62** n = 4,383	0.41** n = 4,294	0.59** n = 4,399	0.56** n = 4,400	1.00 n = 4,616

Table A-2: Average Course Ratings by Division of the Instructor

		Q1		Q2		Q3		Q4		Q5		Q6	
	Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
a	Humanities	119	5.35 [c]	119	5.29	118	4.86	117	5.21 [b]	119	5.28 [b]	119	5.40
b	Sciences	88	5.25	88	5.34	88	4.81	88	5.04 [a]	88	4.94 [a]	88	5.32
c	Social Studies	85	5.15 [a]	85	5.19	85	4.87	85	5.10	85	5.11	85	5.30
	Total	292	5.26	292	5.28	291	4.85	290	5.13	292	5.13	292	5.35
	F		4.39		1.79		0.31		3.43		9.22		1.34
	Sig.		0.01		0.17		0.73		0.03		0.00		0.26

	Eta	Eta^2
Q1	0.17	0.03
Q2	0.11	0.01
Q3	0.05	0.00
Q4	0.15	0.02
Q5	0.24	0.06
Q6	0.10	0.01

Table A-3: Average Course Ratings by Course Level

		Q1		Q2		Q3		Q4		Q5		Q6	
	Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
a	100-Level	114	5.19	114	5.19 [c]	114	4.70 [c]	112	5.00 [c, d]	114	4.99 [c]	114	5.20 [b, c]
b	200-Level	114	5.25	114	5.30	114	4.80 [c]	114	5.14	114	5.14	114	5.37 [a]
c	300-Level	62	5.39	62	5.42 [a]	62	5.17 [a, b]	62	5.26 [a]	62	5.30 [a]	62	5.51 [a]
d	400-Level	12	5.33	12	5.22	11	5.08	12	5.37 [a]	12	5.11	12	5.50
	Total	302	5.26	302	5.28	301	4.85	300	5.13	302	5.12	302	5.34
	F		2.25		2.78		11.75		5.61		4.11		6.79
	Sig.		0.08		0.04		0.00		0.00		0.01		0.00

	R	R^2	Eta	Eta^2
Q1	0.14	0.02	0.15	0.02
Q2	0.13	0.02	0.17	0.03
Q3	0.30	0.09	0.33	0.11
Q4	0.23	0.05	0.23	0.05
Q5	0.17	0.03	0.20	0.04
Q6	0.25	0.06	0.25	0.06

Table A-4: Average Course Ratings by Sex of the Instructor

		Q1		Q2		Q3		Q4		Q5		Q6	
	Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
a	Female	112	5.28	112	5.30	112	4.96	111	5.09	112	5.25	112	5.36
b	Male	184	5.26	184	5.27	183	4.79	183	5.15	184	5.06	184	5.34
	Total	296	5.27	296	5.28	295	4.85	294	5.13	296	5.13	296	5.35
	t		0.41		0.43		2.71		-0.98		2.80		0.43
	Sig.		0.68		0.67		0.01		0.33		0.01		0.67

Table A-5: Average Course Ratings by the Number of Times the Instructor had Taught the Course

		Q1		Q2		Q3		Q4		Q5		Q6	
	Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
a	1 to 2	146	5.20 [c]	146	5.29	145	4.82	146	5.06 [c]	146	5.07	146	5.30
b	3 to 8	84	5.28	84	5.18	84	4.77	83	5.11	84	5.11	84	5.36
c	9 or more	47	5.39 [a]	47	5.37	47	4.97	47	5.26 [a]	47	5.20	47	5.41
	Total	277	5.25	277	5.27	276	4.83	276	5.11	277	5.10	277	5.34
	F		3.07		2.44		2.04		3.06		0.92		0.96
	Sig.		0.05		0.09		0.13		0.05		0.40		0.39

Correlations for the Ungrouped Data

	N	Pearson
Q1	277	0.14*
Q2	277	0.05
Q3	276	0.06
Q4	276	0.12*
Q5	277	0.01
Q6	277	0.08

Table A-6: Average Course Ratings by the Day the Form was Administered

		Q1		Q2		Q3		Q4		Q5		Q6	
	Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
a	Monday	24	5.07	24	5.10	24	4.74	24	4.80 [b, c, d, e]	24	4.85	24	5.11
b	Tuesday	29	5.37	29	5.38	29	4.97	28	5.20 [a]	29	5.12	29	5.41
c	Wednesday	48	5.34	48	5.31	48	4.93	48	5.14 [a]	48	5.14	48	5.40
d	Thursday	54	5.28	54	5.31	54	4.85	54	5.12 [a]	54	5.20	54	5.40
e	Friday	100	5.18	100	5.22	99	4.77	100	5.13 [a]	100	5.03	100	5.28
	Total	255	5.24	255	5.26	254	4.84	254	5.11	255	5.08	255	5.32
	F		2.37		1.41		1.42		3.23		1.85		2.30
	Sig.		0.05		0.23		0.23		0.01		0.12		0.06

	R	R^2	Eta	Eta^2
Q1	-0.04	0.00	0.19	0.04
Q2	-0.01	0.00	0.15	0.02
Q3	-0.07	0.00	0.15	0.02
Q4	0.12	0.01	0.22	0.05
Q5	0.03	0.00	0.17	0.03
Q6	0.01	0.00	0.19	0.04

These data only include courses where the form was completed during the first week of the end-of-course evaluation administration period.

Table A-7: Average Course Ratings by the Time of Day the Form was Administered

		Q1		Q2		Q3		Q4		Q5		Q6	
	Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
a	Before 10 a.m.	87	5.18	87	5.25	87	4.70 [d]	87	5.11	87	5.03	87	5.26
b	10 a.m. to Noon	45	5.25	45	5.26	45	4.84	45	5.11	45	5.03	45	5.33
c	Noon to 3 p.m.	83	5.28	83	5.33	83	4.87	83	5.11	83	5.12	83	5.36
d	3 p.m. to 5 p.m.	34	5.39	34	5.26	33	5.08 [a]	34	5.18	34	5.23	34	5.49
	Total	249	5.26	249	5.28	248	4.83	249	5.12	249	5.09	249	5.34
	F		1.68		0.40		4.31		0.20		1.31		2.14
	Sig.		0.17		0.75		0.01		0.90		0.27		0.10

	R	R^2	Eta	Eta^2
Q1	0.14	0.02	0.14	0.02
Q2	0.04	0.00	0.07	0.00
Q3	0.21	0.05	0.22	0.05
Q4	0.04	0.00	0.05	0.00
Q5	0.12	0.01	0.13	0.02
Q6	0.15	0.02	0.16	0.03

These data only include courses where the form was completed during the first week of the end-of-course evaluation administration period.

Table A-8: Average Course Ratings by Course Type (Instructor-Defined)

		Q1		Q2		Q3		Q4		Q5		Q6	
	Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
a	Mixed	128	5.26	128	5.30	127	4.81 [c]	128	5.12	128	5.17 [b]	128	5.36
b	Lecture	52	5.15	52	5.18	52	4.59 [c, d]	52	5.06	52	4.80 [a, c]	52	5.24
c	Discussion	61	5.29	61	5.30	61	5.03 [a, b]	61	5.22	61	5.33 [b, d]	61	5.42
d	Experiential	41	5.33	41	5.31	41	4.94 [b]	40	5.04	41	5.00 [c]	41	5.30
	Total	282	5.26	282	5.28	281	4.84	281	5.12	282	5.11	282	5.34
	F		1.35		0.83		7.19		1.54		9.62		1.44
	Sig.		0.26		0.48		0.00		0.21		0.00		0.23

	Eta	Eta^2
Q1	0.12	0.01
Q2	0.09	0.01
Q3	0.27	0.07
Q4	0.13	0.02
Q5	0.31	0.09
Q6	0.12	0.02

Table A-9: Average Course Ratings by the Instructor's Tenure Track Status

		Q1		Q2		Q3		Q4		Q5		Q6	
	Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
a	Tenure Track	206	5.29	206	5.31	205	4.86	206	5.18	206	5.17	206	5.40 [c]
b	Not Tenure Track	17	5.23	17	5.20	17	4.85	15	4.98	17	5.20	17	5.27
c	Visiting Scholar	23	5.12	23	5.13	23	4.79	23	4.94	23	5.00	23	5.14 [a]
	Total	246	5.27	246	5.29	245	4.85	244	5.14	246	5.15	246	5.37
	F		1.31		1.57		0.17		3.55		0.99		3.69
	Sig.		0.27		0.21		0.85		0.03		0.37		0.03

	Eta	Eta^2
Q1	0.10	0.01
Q2	0.11	0.01
Q3	0.04	0.00
Q4	0.17	0.03
Q5	0.09	0.01
Q6	0.17	0.03

Table A-10: Average Course Ratings by Instructor's Tenure Status

		Q1		Q2		Q3		Q4		Q5		Q6	
	Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
a	Not Tenured	125	5.23	125	5.23	124	4.82	123	5.07	125	5.15	125	5.32
b	Tenured	122	5.31	122	5.35	122	4.88	122	5.22	122	5.16	122	5.42
	Total	247	5.27	247	5.29	246	4.85	245	5.14	247	5.15	247	5.37
	t		-1.22		-1.83		-0.79		-2.49		-0.15		-1.67
	Sig.		0.22		0.07		0.43		0.01		0.88		0.10

Table A-11: Average Course Ratings by Instructor's Rank

		Q1		Q2		Q3		Q4		Q5		Q6	
	Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
a	Lecturer	26	5.15	26	5.16	26	4.81	24	5.01	26	5.07	26	5.17 [d]
b	Instructor	23	5.21	23	5.29	23	4.87	23	5.01	23	5.26	23	5.30
c	Assistant	113	5.25	113	5.24	112	4.82	113	5.08	113	5.09	113	5.33
d	Associate	65	5.43 [e]	65	5.36	65	4.97	65	5.23	65	5.20	65	5.50 [a]
e	Full	65	5.18 [d]	65	5.31	65	4.78	65	5.18	65	5.08	65	5.32
	Total	292	5.26	292	5.28	291	4.85	290	5.13	292	5.13	292	5.35
	F		2.87		0.96		1.16		2.05		0.86		2.77
	Sig.		0.02		0.43		0.33		0.09		0.49		0.03

	Eta	Eta^2
Q1	0.20	0.04
Q2	0.11	0.01
Q3	0.13	0.02
Q4	0.17	0.03
Q5	0.11	0.01
Q6	0.19	0.04

Table A-12: Average Course Rating Correlations with Instructor's Age and Course Enrollment

		Q1	Q2	Q3	Q4	Q5	Q6
Age of Instructor	Pearson Correlation	-0.01	0.06	0.02	0.19**	0.09	0.08
	Sig.	0.90	0.27	0.68	0.00	0.11	0.19
	N	295	295	294	293	295	295
Course Enrollment	Pearson Correlation	-0.23**	-0.19**	-0.26**	-0.27**	-0.14*	-0.24**
	Sig.	0.00	0.00	0.00	0.00	0.02	0.00
	N	306	306	305	304	306	306

Table A-13: Ratings by Year in College

Student Responses

	Q1		Q2		Q3		Q4		Q5		Q6	
Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
First Year	1,448	5.21	1,371	5.18	1,357	4.73	1,368	5.07	1,376	5.11	1,452	5.29
Sophomore	1,508	5.18	1,454	5.27	1,425	4.83	1,459	5.03	1,457	5.03	1,510	5.27
Junior	800	5.24	760	5.28	739	4.80	779	5.03	769	5.12	802	5.31
Senior	758	5.19	728	5.25	704	4.82	730	4.99	730	5.02	762	5.27
Other	37	5.54	36	5.56	33	4.82	34	5.56	34	5.35	37	5.59
Total	4,551	5.20	4,349	5.24	4,258	4.79	4,370	5.04	4,366	5.07	4,563	5.28

Course-Level Data: Correlations between average course ratings and proportional student composition

	N	% First Year	% Sophomores	% Juniors	% Seniors
Q1	306	-0.08	-0.06	0.07	0.08
Q2	306	-0.12*	0.00	0.13*	0.03
Q3	305	-0.23**	-0.08	0.12*	0.23**
Q4	304	-0.11	-0.04	0.01	0.14*
Q5	306	-0.06	-0.10	0.11*	0.07
Q6	306	-0.17**	-0.10	0.13*	0.18**

Table A-14: Ratings by the Student's Anticipated Grade in the Course

Student Responses

	Q1		Q2		Q3		Q4		Q5		Q6	
Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
A	682	5.28	655	5.43	635	4.82	657	5.23	657	5.18	683	5.43
A-	1,114	5.21	1,068	5.27	1,067	4.83	1,094	5.19	1,073	5.13	1,116	5.36
B+	950	5.27	922	5.28	895	4.84	938	5.09	928	5.14	955	5.38
B	831	5.14	777	5.09	767	4.73	812	4.93	811	4.97	831	5.20
B-	312	5.06	302	5.07	286	4.72	311	4.82	306	4.93	310	5.09
C+	129	5.25	122	5.14	120	4.68	128	4.80	126	4.93	130	5.09
C	144	4.69	135	4.76	136	4.40	140	4.23	138	4.55	145	4.54
D	22	4.82	22	5.00	18	4.06	21	4.33	22	4.86	23	4.61
F	8	4.63	7	5.29	8	4.38	7	4.29	8	4.50	8	4.38
Total	4,192	5.19	4,010	5.23	3,932	4.78	4,108	5.04	4,069	5.06	4,201	5.28

Correlations between student ratings and anticipated grades

	Q1	Q2	Q3	Q4	Q5	Q6
Correlation	0.09**	0.13**	0.07**	0.19**	0.11**	0.17**
N	4,192	4,010	3,932	4,108	4,069	4,201

Course-Level Data: Correlations between average course ratings and proportional composition

	N	% A Grades	% B Grades	% C Grades	% D or F Grades
Q1	303	0.00	0.02	-0.05	-0.09
Q2	303	0.08	-0.05	-0.06	-0.03
Q3	302	0.07	0.00	-0.15*	-0.11
Q4	301	0.11	-0.04	-0.16**	-0.10
Q5	303	0.03	0.03	-0.10	-0.15*
Q6	303	0.00	0.06	-0.13*	-0.09

Table A-15: Ratings by Student's Sex

Student Responses

	Q1		Q2		Q3		Q4		Q5		Q6	
Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
Female	2,566	5.24	2,448	5.28	2,414	4.90	2,459	5.09	2,465	5.14	2,575	5.32
Male	1,902	5.16	1,822	5.20	1,770	4.65	1,832	4.98	1,823	4.99	1,907	5.24
Total	4,468	5.21	4,270	5.24	4,184	4.79	4,291	5.05	4,288	5.07	4,482	5.29

Course-Level Data: Correlations between average course ratings and proportional student composition

	N	% Female	% Male
Q1	305	0.10	-0.10
Q2	305	0.05	-0.05
Q3	304	0.25**	-0.25**
Q4	303	0.08	-0.08
Q5	305	0.20**	-0.20**
Q6	305	0.10	-0.10

Table A-16: Ratings by Anticipated GPA Effect

Student Responses

	Q1		Q2		Q3		Q4		Q5		Q6	
Category	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
Raise GPA	1,317	5.28	1,263	5.35	1,240	4.85	1,286	5.23	1,272	5.16	1,317	5.43
No effect	2,263	5.21	2,155	5.25	2,111	4.82	2,127	5.05	2,152	5.10	2,268	5.28
Lower GPA	800	5.00	762	4.97	742	4.61	791	4.69	777	4.80	804	4.99
Total	4,380	5.20	4,180	5.23	4,093	4.79	4,204	5.03	4,201	5.06	4,389	5.27

Course-Level Data: Correlations between average course ratings and proportional student composition

	N	% Raise GPA	% No Effect	% Lower GPA
Q1	305	-0.05	0.10	-0.06
Q2	305	0.06	0.11*	-0.19**
Q3	304	-0.01	0.15**	-0.15**
Q4	303	0.14*	-0.01	-0.14*
Q5	305	0.01	0.14*	-0.17**
Q6	305	0.06	0.02	-0.09

Table A-17: Ratings by Major Status

Student Responses

	Q1		Q2		Q3		Q4		Q5		Q6	
Course was...	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
In major	1,444	5.25	1,397	5.33	1,377	4.92	1,416	5.16	1,394	5.10	1,443	5.38
In intended major	370	5.25	351	5.23	357	4.78	356	5.18	353	5.14	370	5.43
Not in major	2,021	5.15	1,917	5.20	1,858	4.72	1,919	4.92	1,932	5.01	2,028	5.19
Had not yet declared	669	5.22	637	5.16	621	4.70	637	5.06	642	5.13	674	5.24
Total	4,504	5.20	4,302	5.24	4,213	4.79	4,328	5.04	4,321	5.07	4,515	5.28

Course-Level Data: Correlations between average course ratings and proportional student composition

	N	% In major or intended major	% Not in major or not yet declared
Q1	302	0.12*	-0.09
Q2	302	0.14*	-0.09
Q3	301	0.30**	-0.21**
Q4	300	0.20**	-0.19**
Q5	302	0.05	-0.04
Q6	302	0.23**	-0.19**

Table A-18: The Overall Course Rating as a Function of the Other Items

This material is offered to provide information about the factors affecting the global score (Q6) on the form. Presuming the first five questions on the form were components of the overall score, students implicitly weighted the factors as follows (in descending order of significance): activities during the meeting time (Q1), oral and written work, tests, and other assignments (Q4); course materials (Q5); interactions with the instructor (Q2); interactions with the other students (Q3).

These results come from expressing the mean score for Q6 as a linear function of the mean scores for the other items. In a stepwise fashion, variables were successively entered into the mix to test for explanatory power. The results are listed below.

A note regarding procedure: Some researchers (Studenmund, Anastasi, et al.) have been critical of stepwise procedures because multicollinearity can make it difficult to assess the unique contribution of each independent variable. As such, the order in which items were added to the model could be questioned absent an a priori theoretical underpinning.

Table A-18: The Overall Course Rating as a Function of the Other Items (Continued)

Model Summary (N = 303)

Q6 as a function of...	R^2	R^2 Change	Standard Error of the Estimate
Q1	0.669	0.669	0.270
Q1, Q4	0.732	0.062	0.243
Q1, Q4, Q5	0.758	0.027	0.231
Q1, Q4, Q5, Q2	0.768	0.010	0.227
Q1, Q4, Q5, Q2, Q3	0.772	0.003	0.226

R^2 is an indicator of fit, ranging from zero to one. It can be interpreted as the amount of variance explained by the model.

Coefficients for the Full Model

Variable	Unstandardized Beta Coefficient	Standard Error	Standardized Beta Coefficient	t statistic
(Constant)	0.21	0.17		1.24
Q1	0.42	0.04	0.44	9.76
Q4	0.26	0.04	0.26	7.03
Q5	0.13	0.03	0.16	4.72
Q2	0.12	0.04	0.13	3.21
Q3	0.06	0.03	0.07	2.13

Standardized coefficients are indicators designed to help make the coefficients more comparable with one another.

Correlation Matrix

	Q6 Mean	Q1 Mean	Q2 Mean	Q3 Mean	Q4 Mean	Q5 Mean
Q6 Mean	1.00					
Q1 Mean	0.82	1.00				
Q2 Mean	0.71	0.71	1.00			
Q3 Mean	0.58	0.57	0.53	1.00		
Q4 Mean	0.70	0.61	0.57	0.44	1.00	
Q5 Mean	0.60	0.53	0.51	0.43	0.40	1.00

General Methodological Note:

Measurements obtained with Likert-types of scales are variously viewed as ordinal or interval variables, depending on the project and the practitioner. For a variety of reasons (among them being consistency with the Phase I material) the scales were treated as interval in this report. This may arouse questions regarding measurement scales as well as the treatment of the data as samples versus populations. The parametric techniques employed herein are reasonably robust, and the results garnered could be viewed as samples from a universe of potential repeated trials. Additionally, most of the statistical tests performed were done with course-level data. For those with further interest, the psychometric literature offers considerable research. See also:

Krantz, David, Justifying interval scales, EVALtalk listserv sponsored by the American Evaluation Association.

Ostrom, T.M. & Gannon, K.M., "Exemplar Generation: Assessing How Respondents Give Meaning to Rating Scales," *Methodology for Determining Cognitive and Communicative Processes in Survey Research*, Jossey-Bass, 1996.

Osgood, C., Suci, G., & Tannenbaum, P., *The Measurement of Meaning*, Urbana: University of Illinois Press, 1957.