

AN ANALYSIS OF STUDENT EVALUATIONS OF INSTRUCTION FOR THE FALL QUARTER 2004

RESEARCH NOTES REPORT SERIES

OIART Notes are brief, limited distribution technical reports intended to address a specific question in response to a particular request. These reports are distributed to the office(s) or committee(s) that requested the information and sometimes to other key decision-makers in related areas. The format of these reports is relatively informal and no attempt is made to contextualize the findings within the context of a research literature.

For more information about any of these contact Western Washington University's Office of Institutional Assessment, Research, and Testing (OIART).

Richard Frye, Chris Stark, and Joseph E. Trimble

November, 2005

OFFICE OF INSTITUTIONAL ASSESSMENT, RESEARCH, AND TESTING
WESTERN WASHINGTON UNIVERSITY

RESEARCH NOTES REPORT SERIES • REPORT 2005-01

An Analysis of Student Evaluations of Instruction
Western Washington University
November, 2005

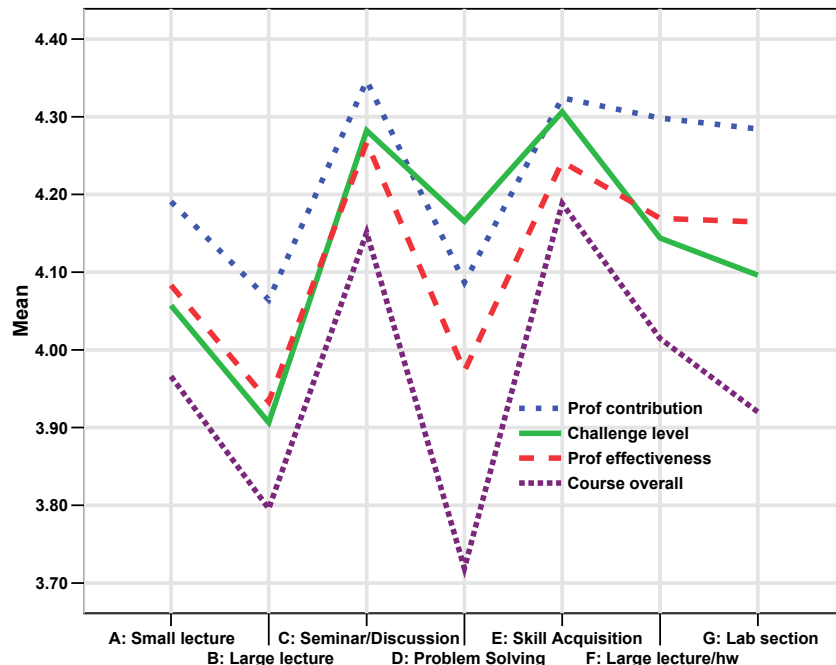
I. Introduction

Over the past three decades, student evaluations of instruction have become the primary documentation of instructor effectiveness, and important inferences are regularly made about courses and instructors on the basis of teaching evaluation ratings. Nevertheless, the instructors and administrators who use them usually have limited understanding of the statistical properties of these ratings which should guide meaningful interpretations, and results are regularly used in ways that are more likely to ensure student *satisfaction* rather than student *achievement*.

This report reviews summary data from all 1175 courses evaluated by the Office of Institutional Assessment, Research, and Testing in Fall quarter, 2004, in comparison with similar data from 2000 and 2002. Because teaching evaluations are not required but voluntary, these courses represent a self-selected, non-random sample of some 64% of the 1,834 courses offered at Western during the quarter, significantly up from 48% of all courses evaluated in Fall 2002. There is no information available about whether and to what extent courses evaluated are different from the courses not evaluated.

Since 1994, Western has offered seven different teaching evaluation forms, each with a different set of questions aimed at the needs of a particular class format. While numerous questions occur on more than one evaluation form, only three questions appear on all forms. These questions ask students to rate the instructor and course on *course overall*, *instructor's teaching effectiveness*, and *instructor's overall contribution*. In addition, each form has a question about the *challenge level* of the class, though exact wording varies somewhat across forms.

Figure 1. Average ratings on common questions, by class format



Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

As shown in Figure 1, ratings on these common questions follow consistent patterns of variation across class formats. In general, *seminars* and *skills acquisition* classes garner the highest ratings on all three summary questions. *Small lecture* classes and *large-lecture-with-homework* classes score lower and comparably to each other, while *problem-solving* classes get lower approval marks, but relatively higher *challenge* ratings. *Large lecture* classes and *labs* typically garner the lowest *course overall* ratings, but labs get higher ratings, more similar to *problem solving* classes, for *instructor effectiveness*, *instructor contribution*, and *challenge level*.

Figure 1 suggests that each class format has a “signature” distribution of ratings which differs substantially in many cases from the overall evaluation mean for each question. The overall mean for each question tends to fall between the “small lecture” mean (31% of courses) and the seminar mean (27% of courses), the most common formats.

These sources of variation are illustrated in Table 1 for one representative item, *instructor effectiveness in teaching this course*, over a four-year period. Course evaluation ratings (see scale below Table 1) are remarkably consistent over time, and show significant and predictable variation due to a number of additional factors not necessarily or directly associated with individual teaching ability, including *challenge level of course*, *expected grade*, *class format*, *reasons for taking the course*, *course level*, *motivation level*, and *general subject area*.

The patterns are similar for *course overall* and *instructor contribution* ratings. In the next section each of these sources of variation is discussed in more detail, along with an exploration of the relative sizes of these effects and their implications for interpreting course evaluation results.

Table 1. Variation in mean scores for "instructor effectiveness" *

Overall mean	<i>Fall 2004 4.05 (median = 4.16; SD = .64); Fall 2000=4.07; Fall 2002= 4.03</i>						
Class format	<i>Seminar</i>	<i>Skills</i>	<i>Small lect</i>	<i>Prob solv</i>	<i>Lab</i>	<i>Lrg lect/hw</i>	<i>Lrg lect</i>
<i>Fall 2000</i>	4.28	4.12	4.05	3.98	3.96	3.85	3.78
<i>Fall 2002</i>	4.17	4.21	4.06	3.95	4.02	4.13	3.86
<i>Fall 2004</i>	4.27	4.24	4.08	3.97	4.16	4.17	3.93
Class reason	<i>Elective</i>	<i>Major</i>	<i>Minor</i>	<i>GUR</i>			
<i>Fall 2000</i>	4.26	4.10	4.08	3.79			
<i>Fall 2002</i>	4.44	4.18	4.37	3.98			
<i>Fall 2004</i>	4.31	4.14	4.18	3.97			
Course level	<i>100</i>	<i>200</i>	<i>300</i>	<i>400</i>	<i>500</i>		
<i>Fall 2000</i>	3.90	4.06	4.00	4.10	4.25		
<i>Fall 2002</i>	4.04	4.15	4.14	4.27	4.38		
<i>Fall 2004</i>	4.01	4.07	4.15	4.21	4.26		
Motivation level	<i>Very low</i>	<i>Low</i>	<i>Moderate</i>	<i>High</i>			
<i>Fall 2000</i>	3.82	3.77	3.92	4.18			
<i>Fall 2002</i>	3.69	3.85	3.97	4.22			
<i>Fall 2004</i>	3.94	4.03	4.19	4.36			
Subject area	<i>Behav.Sci</i>	<i>Educ</i>	<i>Humanities</i>	<i>Soc sci</i>	<i>Science</i>	<i>Engineer'g</i>	<i>Business</i>
<i>Fall 2000</i>	4.24	4.19	4.13	3.98	3.97	3.94	3.76
<i>Fall 2002</i>	4.28	3.98	4.19	4.07	3.95	4.07	3.82
<i>Fall 2004</i>	4.16	4.31	4.15	4.04	4.05	4.27	3.89

* (scale: 5=Excellent, 4=Very Good, 3=Good, 2=Fair, 1=Poor, 0=Very Poor)

Note: Since most evaluation questions are not common to all formats, they do not permit comparison across formats, and will not be analyzed here. However, average ratings for all questions on each form are presented in the Appendix.

II. Analysis

Variation by challenge level of course

All of the Western evaluation forms ask students to rate in some way the level of challenge experienced in the course, on a scale from "excellent" to "very poor." Implicit in this scale of measurement is the idea that there is some "optimal level" of it; there can be either too much or too little, but unfortunately this scale does not permit such distinctions. "Excellent" level of challenge presumably means "just the right amount of stress," one that stimulates engagement and motivates a student's best work. Across all evaluation formats, the average rating on these similar questions is 4.14, about a "B+" on a 5-pt scale, suggesting that in general WWU students feel appropriately challenged by their courses overall.

National data have shown that students give higher course evaluation ratings to more challenging courses, and this is true at Western as well. Students seem to associate "challenge" with having a positive learning experience, and evaluate courses and instruction accordingly. As shown in Figure 1, challenge ratings closely mirror ratings for course and instructor. If, as some faculty believe, good evaluation ratings were awarded to easy-grading instructors, course and instructor ratings would be *inverse* to challenge ratings, and that is not the case. The average challenge ratings are highly correlated with the three common questions (.69-.73), though with smaller variance, and vary across class formats virtually in parallel to the common question ratings.

Challenge level is positively correlated with motivation index (.25, discussed below) and expected grade (.21), and negatively correlated with class size (-.23). Summaries of average ratings on the three common questions, together with the average grade students expected in those courses, are presented in Table 3 below and in Figure 1, above. Comparative data from 2000 and 2002 are included in Table 3 for comparison.

Class format options

Since 1994, Western has offered seven different teaching evaluation forms, each with a different set of questions aimed at the needs of a particular class format, although instructors have been encouraged to use whichever question set will provide them with the most useful feedback about their particular courses. Table 2 shows the distribution of form usage for the Fall 2004 quarter, with considerable variation in class size for each format.

Over several years, about a third of classes each quarter have been *small lecture* classes; about 15% *seminars*; about 15% *skills acquisition* classes; about an eighth (12%) *large lecture* classes (plus another 2% *large lecture with homework*); about 7% *problem solving* classes; and about 6% percent were *labs*. The relatively large standard deviations suggest substantial size variability within each class format.

Comparing figures for 2004 with earlier years suggests at least a temporary shift away from small and large lecture classes toward more seminars and skills acquisition classes, shifts which should improve student engagement if they persist over time, as discussed below.

Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

Table 2. Distribution of class sizes by class format

Class format	% of classes 2000	% of classes 2002	% of classes 2004	Mean class size 2000	Mean class size 2002	Mean class size 2004	Std. Dev class size 2004
A: Small lecture	37.3	35.7	31.1	28	29	29	18.6
B: Large lecture	15.2	16.6	12.1	83	84	88	71.2
C: Seminar/ Discussion	14.4	15.5	26.9	17	20	18	8.6
D: Problem Solving	10.3	9.0	7.0	27	26	25	11.4
E: Skill Acquisition	12.9	12.9	14.8	22	21	21	15.7
F: Large lecture/ hw	2.3	2.7	1.9	62	54	47	33.1
G: Lab section	7.6	7.6	6.2	28	27	24	10.2
Total	100.0	100.00	100.0				

Variation in ratings by class format

In general, each class format has a different profile of average ratings on the three common questions, as shown above in Figure 1, and below in Table 3. *Seminars* and *skills acquisition* classes consistently garner the highest ratings on all three summary questions. *Small lecture* classes and *large-lecture-with-homework* classes score lower and comparably to each other, while *problem solving* classes get lower approval marks but relatively higher *challenge* ratings. *Large lecture* classes and *labs* typically garner the lowest *course overall* ratings, but *labs* get higher ratings, more similar to *problem solving* classes, for *instructor effectiveness*, *instructor contribution*, and *challenge level*. So each class format has a “signature” distribution of ratings which differs substantially in many cases from the overall evaluation mean for each question.

Table 3. Average ratings on common questions, by class format for Fall, 2002 and 2004

Course format	Course overall		Teaching effectiveness		Instructor contribution		Average challenge		Expected grade	
	2002	2004	2002	2004	2002	2004	2002	2004	2002	2004
All formats	3.91	4.00	4.05	4.13	4.16	4.23	4.05	4.14	3.33	3.39
A: Small lecture	3.96	3.96	4.06	4.08	4.18	4.19	4.06	4.06	3.31	3.34
B: Large lecture	3.70	3.79	3.86	3.93	3.98	4.06	3.90	3.91	3.13	3.16
C: Seminar	4.06	4.15	4.18	4.27	4.27	4.35	4.24	4.28	3.56	3.67
D: Problem solving	3.82	3.74	3.96	3.97	4.12	4.10	4.16	4.18	3.26	3.17
E: Skills acquisition	4.13	4.19	4.21	4.24	4.27	4.32	4.29	4.31	3.57	3.50
F: Large lect/ hw	3.93	4.01	4.13	4.17	4.22	4.30	4.07	4.14	3.07	3.14
G: Lab	3.74	3.92	4.02	4.16	4.13	4.28	4.15	4.09	3.24	3.25

* (scale: 5=Excellent, 4=Very Good, 3=Good, 2=Fair, 1=Poor, 0=Very Poor)

This consistent hierarchy in scores suggests that smaller, more interactive formats like seminars or skills acquisition classes consistently earn higher ratings, other things being equal, than larger, less interactive classes like large lectures. These differences tend to persist over time, as shown in Tables 1 and 3, and are consistent with the greater opportunities for student engagement, student-student interaction, and instructor-student interaction possible in these formats.

It is also worth noting that the average expected grade ranges from a “B” in large lectures to an A- in seminars, with an overall average of B+. Students seem to feel they are doing better, or perhaps learning more, in more interactive classes.

Variation among questions within class formats

The nearly parallel lines in Figure 1 indicate that ratings on the common questions are highly correlated with each other, with correlation coefficients in the range of .94 to .97 among *instructor contribution*, *instructor effectiveness*, and *course overall*, and around .7 with *challenge level*. Similarly, all of the other questions on each form are quite highly correlated with the common questions and with each other, generally in the range of .5 to .8. This suggests the likelihood that the various question sets do not discriminate effectively among distinct elements of student course experiences. Psychometric properties of the question sets will be analyzed further in a subsequent report.

The apparent consistency of the relative rankings of (in descending order) *instructor contribution*, *instructor effectiveness*, and *course overall* across class formats is an interesting finding, suggesting that on average students do make some distinctions between the *course* and the *instructor*, fairly consistently giving their highest relative ratings for *instructor effectiveness*, and lowest relative rating to *course overall*. For whatever reason, students seem more critical of courses than of instructors.

Although these numbers mask a great deal of individual variation, nevertheless for about two thirds (68%) of courses evaluated, both *instructor* ratings are higher than the *course overall* rating; for 87% of courses; at least one of the two *instructor* ratings is greater than the *course overall* rating; and in fewer than 7% of courses is the *course overall* rating higher than both *instructor effectiveness* and *instructor contribution*, making it a rather rare event.

Further, in as many as one fifth of classes (19.8%) *course overall* is rated higher than just *instructor effectiveness*; but in only 8% of classes is *course overall* rated higher than just *instructor contribution*.

Courses for which *instructor effectiveness* rating is lower than *course overall* rating tend to have somewhat lower than average ratings on all three questions across class formats, as shown in Table 4, with *seminars* and *large lectures with homework* being notable exceptions. (Excluding seminars from this group lowers the average *course* rating to 3.87, and the *instructor effectiveness* average to 3.72).

A tentative hypothesis is that when the *instructor effectiveness* rating is less than the *course overall* rating, students may be indicating that they wanted more from the instructor than they got. This pattern, when seen across a number of different courses for an individual instructor, might indicate a need for remedial action to improve teaching skills, through Western’s Teaching and Learning Academy or the Center for Instructional Innovation.

Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

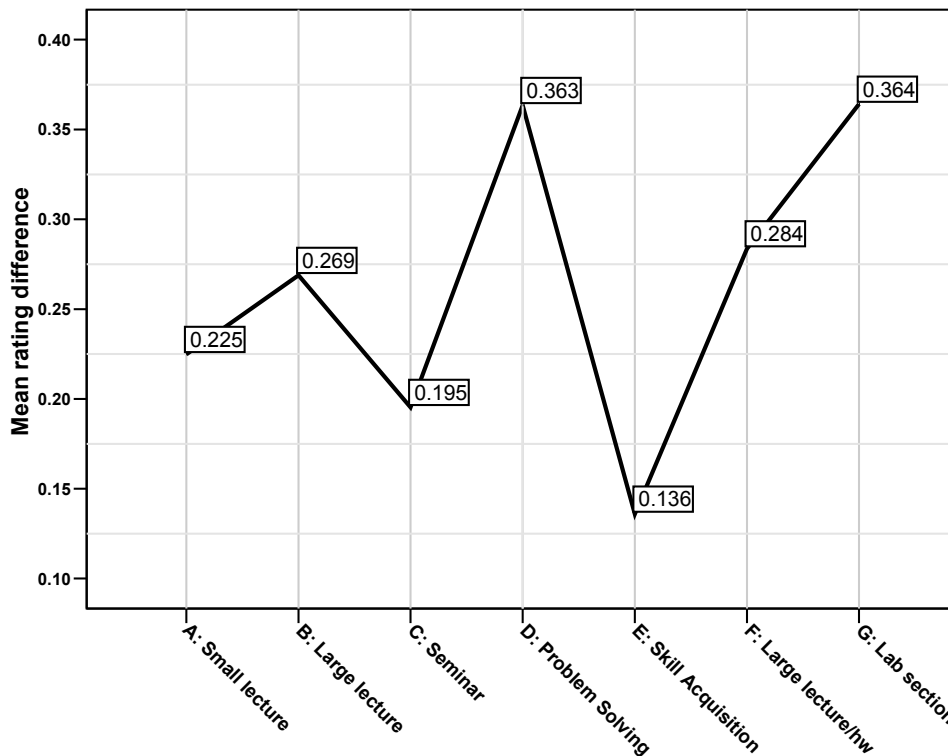
Table 4. Variation in relative ratings when *Course Overall (CO)* rating exceeds *Instructor Contribution (IC)* and/or *Instructor Effectiveness (IE)*

	All courses	CO> IE	CO>IE, Excl seminars	CO > IC	CO>both
Course overall	4.00	3.96	3.87	4.20	4.17
Prof effectiveness	4.13	3.80	3.72	4.02	3.95
Prof contribution	4.23	3.98	3.90	4.02	4.00

While the three questions common to all evaluation forms are very highly correlated, the *differences* between the two *instructor* ratings and the *course* ratings do vary somewhat according to class format. Figure 2 plots the differences between the ratings for *instructor's overall contribution* (generally the highest of the three) and *course overall* (generally the lowest of the three) for each class format.

The differences are largest (diverge the most) for the two most unpopular course formats: *problem solving* courses and *labs*, and smallest (converge the most) for the most popular course formats: *seminars* and *skills* classes. This is an interesting finding, suggesting that students may rate *courses* relatively more harshly than *instructors* in class formats they prefer more (perhaps due to better connection with the instructor in such formats...?), even though in absolute terms they rate both instructors and courses lower in less-preferred class formats.

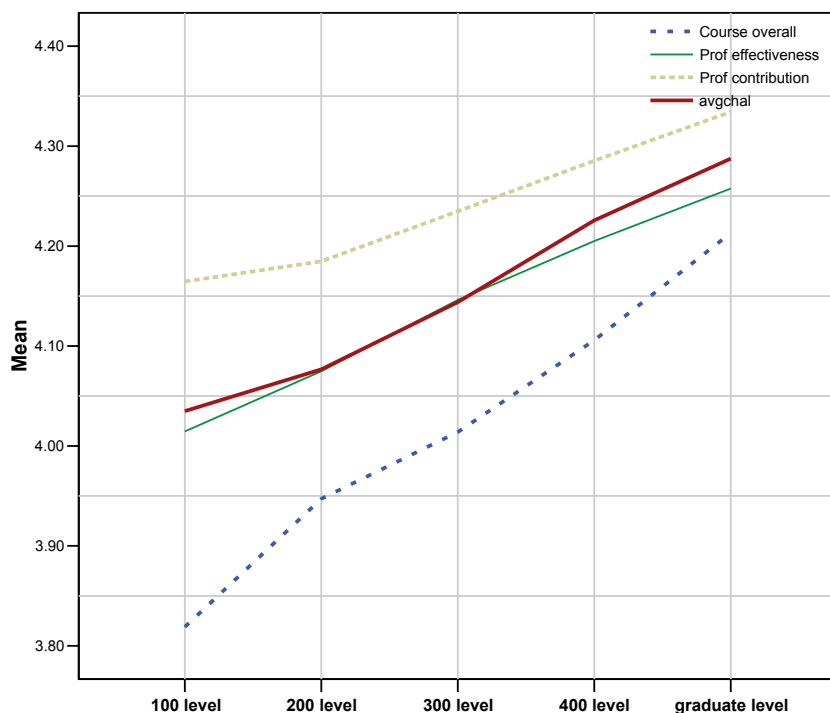
Figure 2. Rating Difference: Prof contribution - Course overall



Variation by course level

As shown in Figure 3 below, and consistent with national findings, student evaluation ratings show significant differences by *course level*. On average, students give lower ratings to 100-level courses than to other class levels, to 200-level courses compared to 300-level, and so on. Senior level (400) courses get consistently higher ratings than lower division courses, and graduate courses regularly garner the highest ratings.

Figure 3. Mean ratings by course level



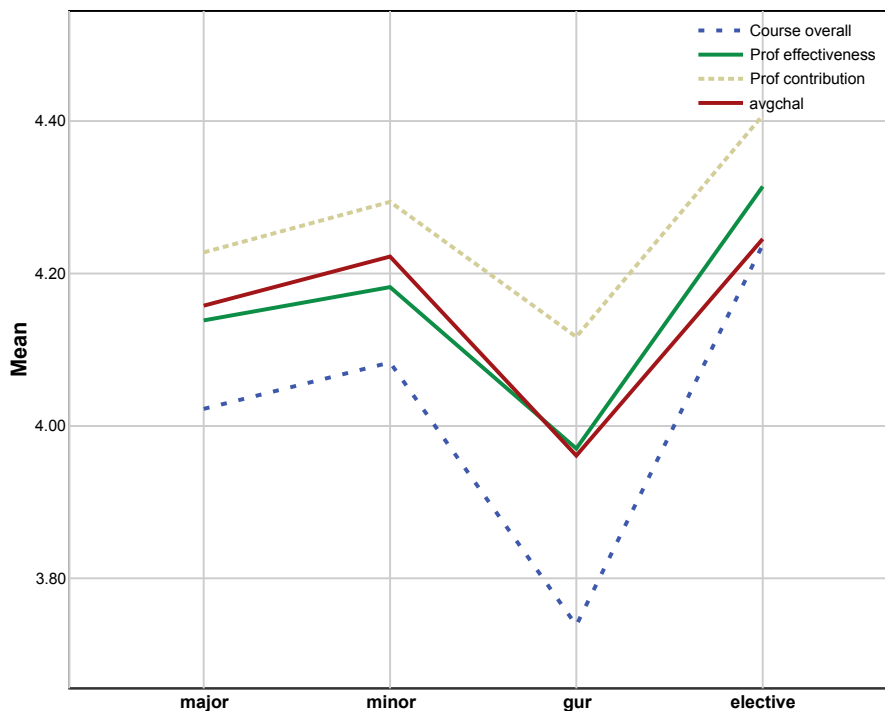
Variation by reason for taking course

All of the evaluation forms ask students why they are taking the course. About two thirds of the courses evaluated (69%, up from 67% in '02) were taken as *major* requirements, about the same as in previous years. About 17% were GUR courses, down from 20% in 2002. About 10% were requirements for a minor (up from 5%), and 5% were electives, about the same as in 2000 and 2002. These changes coincide with a recent reduction in GUR requirements from 72 credits to 60, and may indicate the beginning of a continuing shift away from GUR courses toward more class hours allocated to the major and a minor. It may take several years for a clear pattern to emerge as students adapt to the new general education requirements.

Student ratings on the common questions continued to show significant variation in 2004 according to the student's primary *reason for taking the course*. As shown in Figure 4, electives and courses in a minor earned significantly higher ratings than courses required for the major, which in turn earned much higher ratings than general education requirements.

These findings are consistent with national findings suggesting that required courses may receive lower ratings primarily because they are of less interest to students than major courses, and electives are by their nature usually subjects of particular interest to students. Inclusion on evaluations of an item to assess, with more granularity, student interest in taking a course (*how much* they wanted to take the course, e.g.) might prove informative.

Figure 4. Mean ratings by reason for taking class

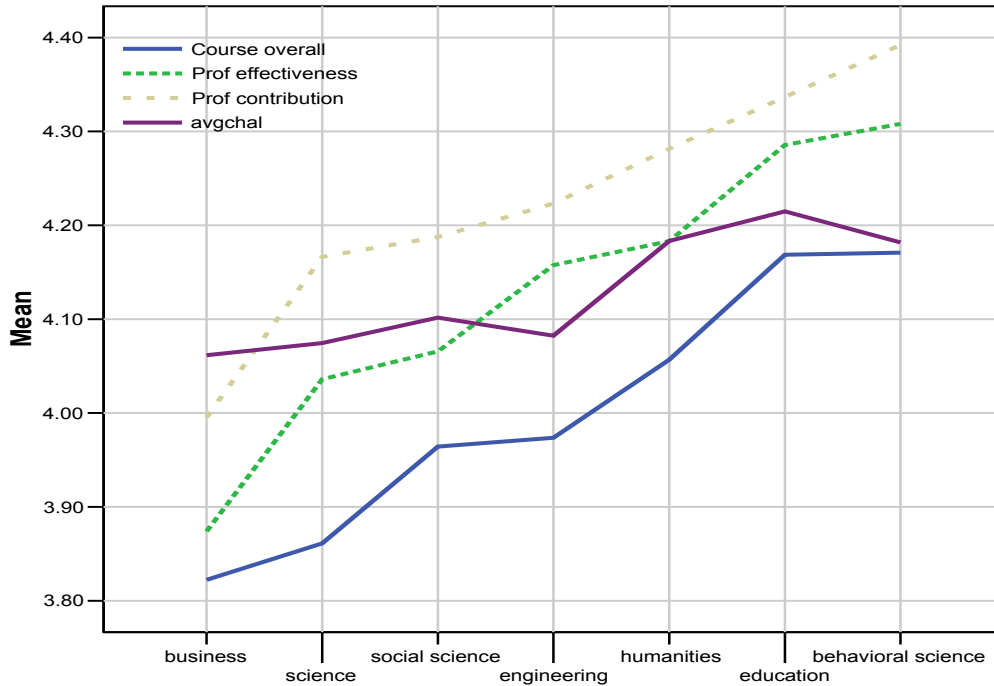


Variation by general subject area

Evaluation ratings also vary significantly by *subject area*, as shown in Figure 5. Both at Western and around the nation, courses in behavioral science, education, and the humanities consistently get the highest ratings, followed by a second group of courses in social sciences and engineering, and then a third group which includes physical sciences and business courses. This clustering of ratings by field is consistent with national findings, with the minor difference that nationally, business course rating averages more often fall between social and physical sciences, with physical sciences usually getting the lowest ratings.

Such lower ratings in the sciences on the national level may be related to a tendency in recent years for such programs to overload students with increasing amounts of course material, as well as by an ever-increasing number of required courses to complete some programs. Reasons for other differences across subject areas are more elusive; it is not known if these differences represent variations in the quality of teaching, the nature of the material, the nature of the students or faculty drawn to the different subject areas, some other factor, or some combination. What is clear at Western is that these rankings are relatively stable over time.

Figure 5. Mean ratings by general subject area



One plausible explanation may be that different subject areas use different proportions of more and less popular class formats. For example, only 7% of business courses and 8% of science courses are *seminars* or *skills* courses (the highest rated formats), compared with 27% in social sciences, around 40% in humanities, behavioral science, and engineering, and 57% in education. To the extent that student ratings are a surrogate for learning, all subject areas might benefit by shifting courses to more interactive class formats whenever possible.

Motivation Index

A "motivation index" was constructed from student responses to the question, "Was this a course you wanted to take?" The index was computed for each course as the difference between the number of "yes" and "no" responses divided by the total number of all responses (*yes, no, or neutral*). The resulting index had a maximum possible range of plus one (all "yes") or minus one (all "no"). The mean value over all courses was .68, with a standard deviation of about .26.

Ratings on the three common questions show modest but significant correlations with the motivation index, between .24 (*instructor contribution*) and .37 (*course overall*); the higher the index, the higher the ratings.

As shown in Table 5 and Figure 6, in Fall quarter 2004, courses in the highest quartile of motivation indices had evaluation scores on the three common questions in the range from 4.31

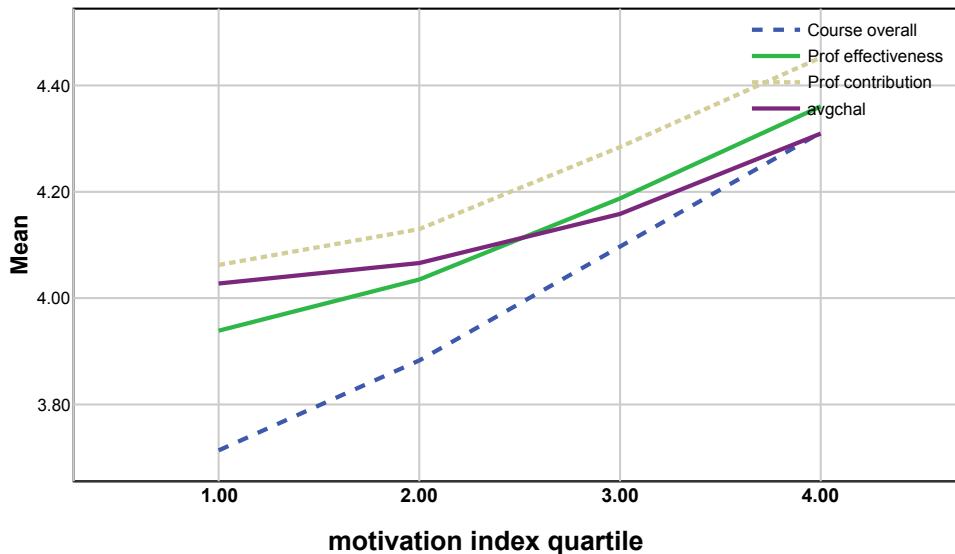
(course overall) to 4.45 (instructor contribution), while the 25% with lowest motivation indices had average scores on the same questions in the range of 3.7 to 4.06. Clearly, Western students give higher ratings to courses they are more motivated to take.

Table 5. Ratings by motivation index quartile* and year

Motivation quartile		Lowest	Low	High	Highest
Course overall	2004	3.71	3.88	4.10	4.31
	2002	3.55	3.86	3.99	4.26
	2000	3.60	3.74	3.99	4.16
Prof effectiveness	2004	3.94	4.03	4.19	4.36
	2002	3.85	4.00	4.11	4.31
	2000	3.81	3.89	4.11	4.24
Prof contribution	2004	4.06	4.13	4.29	4.45
	2002	3.94	4.12	4.21	4.40
	2000	3.96	4.02	4.21	4.31
Challenge level	2004	4.03	4.07	4.16	4.31
	2002	3.96	4.07	4.13	4.27
	2000	4.00	4.03	4.14	4.25

*Lowest qtr, mid-low qtr, mid-high qtr, highest qtr

Figure 6. Mean ratings by motivation level of students



Variation with expected grade

All evaluation forms ask student what grade they expect in the course. As shown in Table 3 above, *expected grades* vary significantly by class format in proportion to *course overall* ratings, with modest positive correlations between *expected grade* and *instructor* ratings, in the range of .29 to .39. Does this mean either that higher ratings are associated with improved learning, or that students “reward” courses in which they expect a better (or “easier”) grade?

Since, as noted above, more challenging courses are rated higher than less challenging courses, the commonly accepted view in national research is that students in fact reward the

perception of better learning with better evaluations. The consistently higher ratings that some class formats elicit may be related to students' perceptions of greater learning in those formats, through higher levels of interaction with faculty and other students, or via the better engagement opportunities offered by some course formats.

Because there is some controversy over the correlation between evaluation ratings and student grades, it is useful to explore the relationships among expected grades, actual grades earned, and course and instructor evaluation ratings.

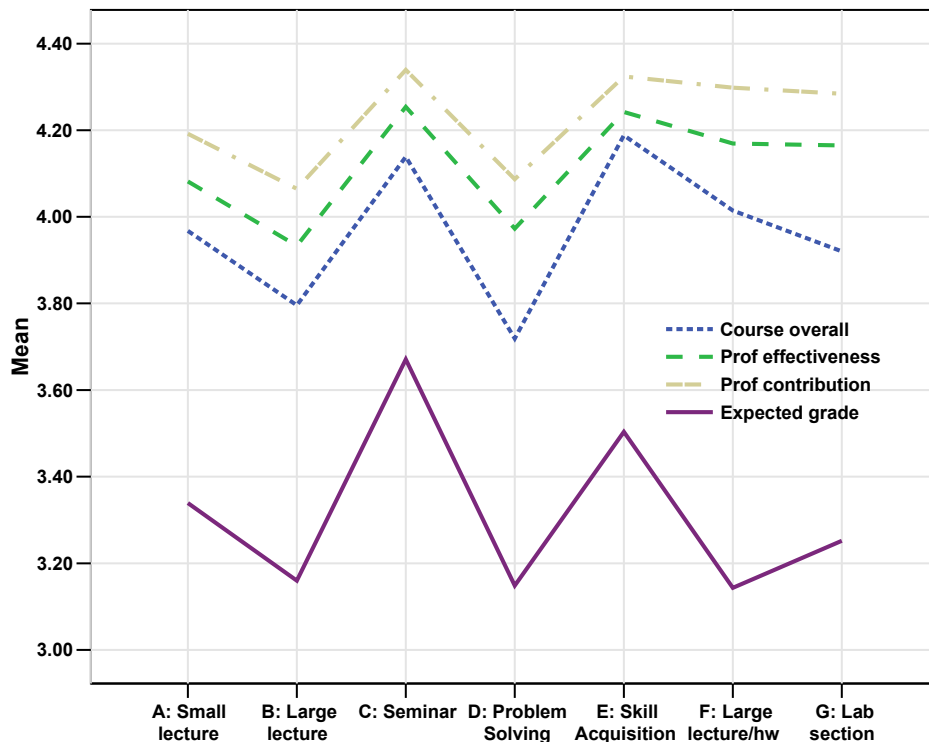
Expected grade and actual grade

A randomly selected sample of 200 courses from Fall quarter 2000 was selected for an exploratory analysis. Results showed that on average students expected higher grades (3.3) than they actually received (3.1), and that the correlation between expected grade and actual grade received was high, about .83 overall, and quite consistent across all ratings. Regression analysis confirmed a strongly significant linear relationship ($p < .000$) between expected grade and grade actually received across all evaluations, suggesting that average class grades can be somewhat reliably predicted from average expected grade:

$$\text{actual final grade} = (1.1 \times \text{expected grade}) - .5$$

This equation implies that student expectations are in greater error for lower grades; the disparity between expected grade and actual grade decreases as grades get higher. On average, students who expect a 2.0 actually get a 1.7; those expecting a 3.0 actually get a 2.8; those expecting a 4.0 get a 3.9. Nationally, students generally benefit from the more frequent feedback about their progress toward course goals provided by more frequent graded assignments; perhaps they would also be able to predict their actual grades better as well.

Figure 7. Evaluation ratings and expected grade



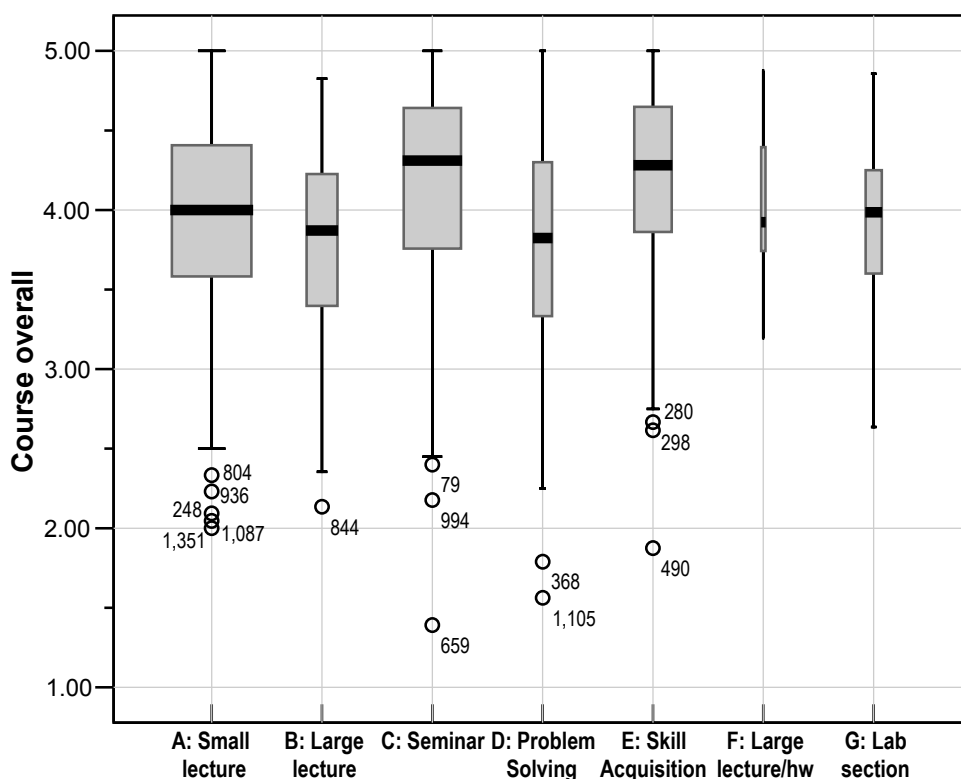
III. Interpreting Evaluation Results

Course overall rating

As discussed in section 2, course evaluation ratings co-vary with a number of independent variables; it would be valuable to know how much of the variation in evaluation ratings is associated with actual differences in instructor proficiency, and how much is associated with other variables not under the instructor's control.

A useful view of the variation in *course overall* ratings is shown the boxplot in Figure 8. The box above each class format is centered about the median rating for that format; each box extends between the 25th percentile and the 75th percentile, and its width represents the proportion of all evaluations in that class format. The "legs" above and below each box show the range of the highest and lowest quartiles, respectively. The few outlying ratings are plotted individually.

Figure 8. Distribution of *Course Overall* ratings



It is clear from the chart that all of the distributions are centered between about 3.7 (*problem solving*) and 4.3 (*skills*), in the general range of “*very good*” to “*excellent*,” confirming the notion that the “default” rating is on average quite high. If we are to take students at their word, the implication is that they are very satisfied with the quality of instruction at Western. While we can safely conclude that the highest-rated courses and instructors are acceptably competent, we do not know whether the lowest-scoring courses and professors are “competent enough.”

Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

The long lower “tails” of each distribution demonstrate visually a considerable degree of skewness in the distributions; ratings are concentrated at the top of each distribution, and extend over a longer range at the lower end, including a small number of “outlier” ratings considerably below the rest.

Indeed, the structure of the evaluations ratings, with 5=*excellent*, 4=*very good*, 3=*good*, 2=*fair*, 1=*poor*, and 0=*very poor*, sets the mark for “poor” all the way down to 1.0, while the lower tail of each distribution extends essentially between 2.5 and 3.5. Without some additional qualitative data, such as interviews with students, it is not possible to link these *relative* numbers to *absolute* measures of teaching ability, student satisfaction with courses, or student learning. That is, we don’t know if there is a cutoff score for either instructor or course ratings above which is “acceptable” and below which is “unacceptable.” We can, however, look more closely at the external variables discussed in Section 2 to learn more about the relative magnitudes of their effects on overall ratings.

Table 5 shows the results of a stepwise regression of the independent variables discussed in section 2 on the *course overall* ratings. By itself *average challenge* level of the course accounts for about half the variance(53%) in *course overall* rating. *Average expected grade* adds another 5%; *motivation index* adds another 3%; *class size* adds another 2%; and *class format*, while statistically significant with such a large sample, add negligibly to the information already in the model. Overall the model “explains” 63% of the variation in the *course overall* ratings across all class formats.

Table 5. Model Summary, Course Overall

Model	R	Incremental R ²	Cumulative R ²	Std. Error
1 (Constant), challenge	.73	.53	.53	.43
2 (Constant), challenge ,expected grade	.77	.05	.58	.40
3 (Constant), challenge ,expected grade, motivation index	.78	.03	.61	.39
4 (Constant), challenge ,expected grade, motivation index, class size	.79	.02	.63	.38
5 : (Constant), challenge ,expected grade, motivation index, class size, class format	.79	.00	.63	.38

We also know that *challenge level* is by itself highly correlated with *course overall* ($r=.73$), *instructor effectiveness* (.69), and *instructor contribution* (.69); in a sense the four variables appear to be measuring only slightly different aspects of some overall measure of teaching ability, student satisfaction with courses, or student learning. If we remove *challenge level* from the analysis, the *total* variation explained by the remaining variables drops substantially, but the *proportion* explained by the other variables increases, as shown in Table 6. The proportion of variance explained by *expected grade* increases from 3% to 14.5% when challenge level is left out of the equation, and the revised model explains about a quarter of the overall variation in *course overall*.

Table 6. Model Summary, Course Overall, excluding challenge level

Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

Model	R	Incremental R ²	Cumulative R ²	Std. Error
1(Constant), expected grade	.381	.145	.144	.57158
2 (Constant), expected grade, motivation index	.483	.087	.232	.54142
3 (Constant), expected grade, motivation index, class format	.486	.03	.235	.54061

Instructor Ratings

The boxplot in Figure 9 shows the distribution of ratings for *instructor contribution*; the distribution of ratings for *instructor effectiveness*, shown in Appendix A, is very similar to *course overall* and *instructor contribution*, but lies between them, as shown above in Figure 1.

Figure 9. Distribution of *Instructor Contribution* ratings

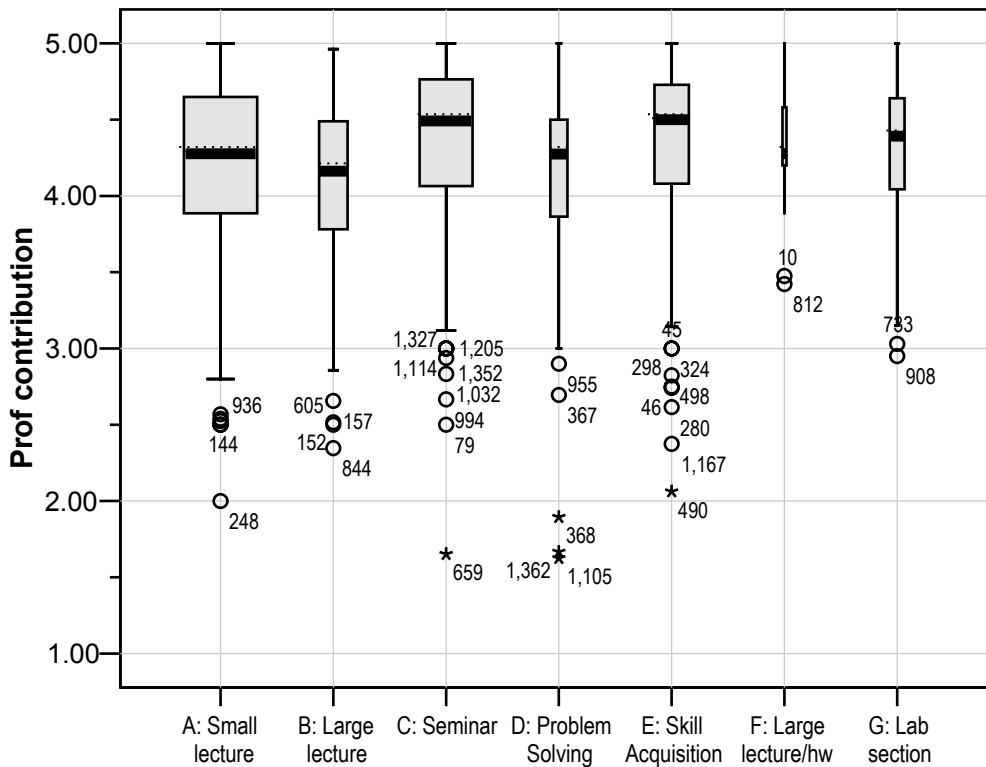


Table 7 shows that *average challenge* level accounts for nearly half the variation (48%) in *instructor contribution*; *average expected grade* adds another 2%; *motivation index* adds another 1%; and *class size* adds 1%. Overall the model “explains” 53% of the variation in the *instructor contribution* ratings across all class formats.

Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

Table 7. Model Summary, *Instructor contribution*

Model	R	Incremental R ²	Cumulative R ²	Std. Error
1 (Constant), challenge	.69 --	.48	.48	.41
2 (Constant), challenge ,exp grade	.71	.02	.50	.41
3 (Constant), challenge ,exp grade, class size	.72	.01	.51	.40
4 (Constant), challenge ,exp grade, class size, motivation index	.72	.01	.52	.40
5 (Constant), challenge ,exp grade, class size, motivation index, reason	.72	.00	.52	.40

Excluding *challenge level* from the analysis yields the model summarized in Table 8, with about 12% of the variation in *instructor contribution* explained by *expected grade*, *class format*, and *motivation index*: the same variables that explained 23% of variation in *course overall* explain only about 12% of the variation in *instructor contribution*.

Table 8. Model Summary *Instructor contribution (challenge excluded)*

Model	R	Incremental R ²	Cumulative R ²	Std. Error
(Constant), avexpgrd	.279	.078	.078	.55
(Constant), avexpgrd, motivation index	.334	.32	.110	.54
(Constant), avexpgrd, motivation index, class format	.343	.05	.115	.54

The high correlations among student ratings for *challenge*, *course overall*, *instructor effectiveness*, and *instructor contribution* suggest that while students make some distinctions among the different questions, the four common questions really measure a rather undifferentiated, composite satisfaction with their entire experience of the course-and-instructor gestalt.

Nevertheless, students do appear to make some distinctions between the instructor and the course; roughly speaking, it appears that about 75% of the variation in *course* satisfaction and nearly 90% of the variation in *instruction* satisfaction are independent of differences in “situational” variables. Further, given the high correlations among the four common measures, a tentative hypothesis is that taken together, the four measures say something about a student’s sense of engaged learning in a class.

The data suggests that students measure their own learning partly by the grade they expect to receive; they assess their level of engagement by their interest in the subject, the class format and size, and class level; and they assess the quality of their educational experience by a combination of challenge level, course relevance, and instructor performance.

Summary

Analysis of course evaluations for Fall 2004 confirms, as in 2000 and 2002, significant and persistent variation in course and instructor ratings associated with many factors *other* than the quality of teaching of individual instructors. Therefore, when using evaluations to make inferences about teaching ability or course quality, it is essential to consider the specific context of each course.

Students give consistently higher ratings to seminars over other formats, to upper division courses over lower division courses, to electives and major courses over GUR's, to "soft" subjects over technical subjects, to courses they want to take over courses they are required to take, and to courses in which they feel they are doing well over courses where they are not.

As shown in Table 1 above, evaluation ratings are sensitive to differences in *class format*, *course level*, *reason for taking the course*, *student motivation level*, *student sense of learning* (as indicated by expected grade), and *general subject area*. Even if teaching skills were equal across all instructors, we would expect the lowest ratings to occur in large, required, GUR lecture courses of a technical nature in business, engineering, or science, and the highest ratings to occur in a small senior or graduate elective seminar in humanities, education, or behavioral science.

One fairly clear and not very surprising conclusion is that students do in fact give higher ratings to formats which promote interaction with instructor and other students, in subjects which are interesting and relevant to their lives and careers, and which best foster a sense of learning and involvement. These factors have all been identified nationally as general "best practices in teaching" to improve student learning; therefore, it can tentatively be concluded that student evaluations of courses and instruction do say something meaningful about student perception of their own learning.

Given that distributions of evaluation ratings are skewed toward the high end, especially for some questions and some class formats, that instructors who choose to have courses evaluated may be different from those who do not, and that the general population of instructors is probably quite skilled, it is not at all clear what *relative rankings* imply for the *absolute ability* of individual instructors. Further, some low evaluation ratings might just indicate a poor fit between instructor skills and course assignment. For example, some instructors may be much better at teaching seminars than lectures, or vice versa. Finally, studies elsewhere have even demonstrated that physically attractive or otherwise charismatic instructors get better ratings than unattractive or more introverted ones.

Because of the number and magnitude of these confounding factors, course evaluation ratings are not a particularly reliable tool for measuring "teaching ability," and should be heavily supplemented with additional data from peer evaluations, course portfolios, teaching portfolios, self-assessments, or other independent instruments. However, student evaluations are helpful for identifying patterns of consistent strengths or weaknesses across many courses; instructors who would like to improve their teaching will find many helpful resources at the Center for Instructional Innovation located in Miller Hall 156.

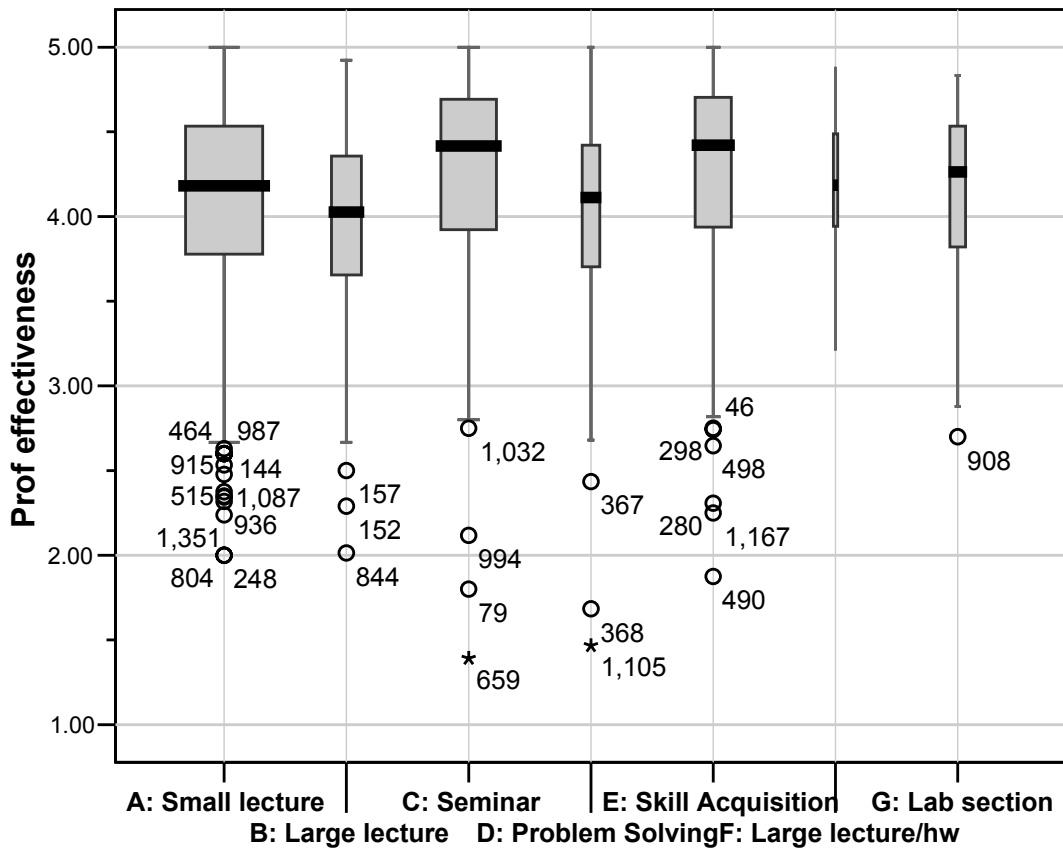
Further analysis is planned to investigate in more detail the statistical properties of evaluation question sets, and to explore options for improving those properties.

Appendix A: Distribution of *Instructor Effectiveness* ratings

Appendix B: Average ratings of teaching evaluations by class format

Appendix A

Figure A-1. Distribution of *Instructor Effectiveness*



ratings

Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

Appendix B. Average ratings of teaching evaluations by class format

Form A: Small Lecture Class	2000	2002	2004
1. Clarity of course goals and objectives was:	3.98	3.99	4.04
2. Challenge level of assigned work was:	4.07	4.06	4.06
3. Fairness of evaluation procedures was:	3.94	3.96	4.03
4. Intellectual challenge offered by the course was:	4.10	4.12	4.11
5. Organization of the course was:	3.91	3.90	3.93
6. Instructor's classroom presentation was:	3.98	4.00	3.99
7. Instructor's use of classroom time was:	4.00	4.01	3.99
8. Instructor's answers to students questions were:	4.04	4.05	4.11
9. Instructors explanations were:	4.02	4.01	4.08
10. Instructor's use of examples was:	4.13	4.13	4.15
11. Instructor's availability for extra help was:	4.06	4.10	4.13
12. Instructor's enthusiasm for the subject was:	4.46	4.49	4.51
13. Instructor's prompt response to homework/tests was:	4.11	4.14	4.14
14. Instructor's record for coming to class on time was:	4.54	4.52	4.53
15. Instructor's record for meeting as scheduled was:	4.56	4.58	4.58
16. Instructor's lecture pace was:	3.86	3.87	3.89
17. Instructor's awareness of student comprehension was:	3.71	3.71	3.76
18. The course overall was:	3.93	3.93	3.97
19. Instructor's effectiveness in teaching subject was:	4.05	4.06	4.08
20. Instructor's contribution overall to the course was:	4.16	4.18	4.19
Course level	3.06	3.03	3.12
Challenge level	4.07	4.06	4.06
Motivation index	0.76	0.73	.72
Expected grade	3.31	3.31	3.34
	(n=354)	(n=362)	(n=379)

Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

<i>(questions in red below are different from Form A)</i>			
Form B: Large Lecture Class	2000	2002	2004
1. Clarity of student responsibilities and requirements	3.91	3.92	3.98
2. Organization of the course	3.79	3.92	3.85
3. Challenge level of assigned work	3.83	3.92	3.83
4. Fairness of evaluation procedures	3.78	3.76	3.91
5. Intellectual challenge offered by course	3.88	3.92	3.93
6. Instructor's classroom presentation	3.74	3.79	3.87
7. Instructor's answers to student questions	3.81	3.83	3.95
8. Instructor's ability to provide alternative explanations	3.77	3.81	3.90
9. Instructor's use of examples and illustrations	3.94	4.03	4.09
10. Instructor's availability for extra help	3.82	3.82	3.88
11. Instructor's enthusiasm for the subject	4.34	4.40	4.47
12. Instructor's record for coming to class on time	4.45	4.38	4.52
13. Instructor's record for meeting as scheduled	4.54	4.50	4.56
14. Instructor's use of class time	3.99	4.01	4.07
15. Instructor's lecture pace	3.60	3.65	3.73
16. Instructor's awareness of student comprehension	3.40	3.43	3.55
17. Instructor's exam questions relative to lectures	3.51	3.58	3.68
18. The course overall	3.67	3.70	3.80
19. Instructor's effectiveness	3.78	3.86	3.93
20. Instructor's contribution overall	3.92	3.98	4.06
Course level	2.11	2.18	1.98
Challenge level	3.91	3.91	3.91
Motivation index	0.66	0.66	.63
Expected grade	3.09	3.13	3.16
	<i>(n=146)</i>	<i>(n=168)</i>	<i>(n=149)</i>

Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

<i>(questions in red below are different from Form A)</i>			
Form C: Seminar or Discussion Group	2000	2002	2004
1. Use of class time was:	4.04	3.92	4.02
2. Clarity of student responsibilities and requirements	4.06	3.89	3.96
3. Encouragement of student self-expression	4.42	4.33	4.34
4. Conduciveness of class atmosphere to student learning	4.19	4.10	4.18
5. Challenge level of assigned work	4.15	4.08	4.15
6. Fairness of evaluation procedures	4.17	4.00	4.13
7. Intellectual challenge offered by the course	4.25	4.12	4.19
8. Relevancy of course content in terms of the field	4.40	4.29	4.35
9. Instructor's preparation for class	4.36	4.23	4.36
10. Instructor's guidance as a discussion leader	4.25	4.13	4.24
11. Instructor's contribution to the discussion	4.36	4.27	4.36
12. Instructor's use of questions/problems	4.23	4.12	4.22
13. Instructor's openness to student views	4.39	4.29	4.34
14. Instructor's enthusiasm for the subject	4.61	4.60	4.60
15. Instructor's record for coming to class on time	4.62	4.49	4.63
16. Instructor's record for meeting as scheduled	4.67	4.57	4.64
17. Instructor's support for student/teacher partnership in learning	4.42	4.32	4.38
18. The course overall was:	4.20	4.06	4.15
19. Instructor's effectiveness in teaching the subject matter	4.28	4.17	4.27
20. Instructor's contribution overall to the course was:	4.37	4.27	4.35
Course level	3.66	3.51	3.69
Challenge level	4.29	4.24	4.28
Motivation index	0.73	0.75	.67
Expected grade	3.61	3.56	3.67
	(n=135)	(n=159)	(n=266)

Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

<i>(questions in red below are different from Form A)</i>			
Form D: Problem Solving Class	2000	2002	2004
1. Instructor's use of classroom time was:	3.75	3.76	3.87
2. Organization of the course was:	3.76	3.79	3.84
3. Contribution of assignments to understanding course content:	3.89	3.88	3.85
4. Clarity of student responsibilities and requirements was:	3.86	3.79	3.81
5. Challenge level of assigned work was:	3.99	4.04	4.03
6. Fairness of evaluation procedures was:	3.91	3.89	3.88
7. Intellectual challenge offered by the course was:	3.97	4.04	4.05
8. Instructor's explanations were:	3.80	3.74	3.78
9. Instructor's ability to provide alternative explanations:	3.86	3.83	3.84
10. Instructor's use of examples and illustrations was:	3.97	3.93	3.96
11. Instructor's ability to deal with student difficulties was:	3.88	3.88	3.88
12. Instructor's answers to student questions were:	3.92	3.89	3.92
13. Instructor's availability for extra help was:	4.10	4.14	4.13
14. Instructor's enthusiasm for the subject was:	4.38	4.37	4.37
15. Instructor's record for coming to class on time was:	4.51	4.59	4.51
16. Instructor's record for meeting as scheduled was:	4.60	4.68	4.59
17. Instructor's prompt response to homework was:	4.08	4.05	4.05
18. The course overall was:	3.78	3.82	3.74
19. Instructor's effectiveness in teaching the subject:	3.98	3.96	3.97
20. Instructor's contribution overall to the course was:	4.12	4.11	4.10
Course level	2.11	2.19	2.23
Challenge level	4.17	4.16	4.18
Motivation index	0.63	0.61	.51
Expected grade	3.19	3.26	3.17
	(n=98)	(n=91)	(n=86)

Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

<i>(questions in red below are different from Form A)</i>			
Form E: Skills Acquisition Class	2000	2002	2004
1. Use of class time was:	4.01	4.08	4.14
2. Sequential development of skills was:	3.93	4.00	4.11
3. Demonstrations of expected skills were:	3.95	4.02	4.10
4. Opportunities for practicing what was learned were:	4.13	4.14	4.25
5. Clarity of student responsibilities and requirements was:	3.95	4.02	4.08
6. Challenge level of assigned work was:	4.12	4.11	4.16
7. Fairness of evaluation procedures was:	4.12	4.14	4.15
8. Instructor's preparation for class was:	4.27	4.29	4.36
9. Instructor's ability to deal with student difficulties was:	4.03	4.11	4.18
10. Instructor's recognition of student progress was:	3.92	4.04	4.07
11. Instructor's availability for extra help was:	4.04	4.12	4.20
12. Instructor's tailoring of instruction to varying skill levels was:	3.81	3.99	4.00
13. Instructor's record for coming to class on time was:	4.48	4.47	4.61
14. Instructor's record for meeting with the class as scheduled was:	4.51	4.58	4.64
15. Instructor's feedback regarding skill performance was:	4.00	4.08	4.10
16. Instructor's monitoring of skill acquisition was:	3.84	3.99	4.02
17. Instructor's ability to break skills into meaningful components was:	3.95	4.08	4.12
18. The course overall was:	4.03	4.13	4.19
19. Instructor's effectiveness in teaching the subject matter was:	4.13	4.21	4.24
20. Instructor's contribution overall to the course was:	4.19	4.27	4.32
Course level	2.66	2.54	2.66
Challenge level	4.29	4.29	4.31
Motivation index	0.88	0.85	.81
Expected grade	3.50	3.57	3.50
	(n=122)	(n=128)	(n=180)

Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

<i>(questions in red below are different from Form A)</i>			
Form F: Large Lecture/Homework	2000	2002	2004
1. Organization of the course was:	3.91	4.13	4.12
2. Opportunity for questions was:	4.08	4.13	4.30
3. Usefulness of course content was:	3.58	3.92	4.01
4. Challenge level of assigned work was:	3.89	3.98	4.03
5. Relationship of exams to emphasized material :	3.77	3.96	4.08
6. Fairness of evaluation procedures was:	3.77	3.84	4.11
7. Instructor's preparation for class was:	4.17	4.33	4.35
8. Instructor's use of examples and illustrations was:	4.01	4.18	4.27
9. Instructors explanations were:	3.75	4.03	4.05
10. Instructor's answers to student questions were:	3.86	4.03	4.11
11. Instructor's ability to deal with student difficulties was:	3.69	3.82	4.04
12. Instructor's availability for extra help was:	3.82	3.98	4.18
13. Instructor's enthusiasm for the subject was:	4.40	4.52	4.52
14. Instructor's ability to make clear concepts and ideas was:	3.69	4.01	4.05
15. Instructor's record for coming to class on time was:	4.54	4.47	4.41
16. Instructor's record for meeting with the class as scheduled was:	4.60	4.52	4.63
17. Instructor's promptness in returning homework was:	4.00	4.05	4.16
18. The course overall was:	3.68	3.93	4.01
19. Instructor's effectiveness in teaching the subject matter was:	3.85	4.13	4.17
20. Instructor's contribution overall to the course was:	4.05	4.22	4.30
Course level	1.91	2.39	2.50
Challenge level	4.03	4.07	4.14
Motivation index	0.49	0.61	.63
Expected grade	2.96	3.07	3.14
	(n=22)	(n=28)	(n=21)

Analysis of Student Evaluations of Instruction
Office of Institutional Assessment, Research, and Testing

<i>(questions in red below are different from Form A)</i>			
Form G: Lab Section	2000	2002	2004
1. Clarity of lab section assignments was:	3.66	3.68	3.95
2. Use of lab section time was:	3.83	3.79	4.05
3. Implementation of safety procedures was:	4.16	4.12	4.30
4. Usefulness of lab section content was:	3.63	3.74	3.92
5. Clarity of student responsibilities and requirements was:	3.75	3.73	3.94
6. Coordination between lectures and lab activities was:	3.27	3.30	3.46
7. Challenge level of assigned work in lab section was:	3.69	3.73	3.81
8. Fairness of evaluation procedures used for lab section was:	3.94	3.86	4.07
9. Lab Instructor's preparation for lab sessions was:	4.13	4.19	4.36
10. Lab Instructor's use of questions/problems was:	3.95	4.03	4.16
11. Lab Instructor's ability to deal with student difficulties was:	4.05	4.06	4.20
12. Lab Instructor's answers to student questions were:	4.03	4.05	4.21
13. Lab Instructor's availability for extra help was:	3.97	4.03	4.14
14. Lab Instructor's enthusiasm for the subject was:	4.23	4.28	4.39
15. Lab Instructor's record for coming to class on time was:	4.59	4.63	4.71
16. Lab Instructor's record for meeting with the class as scheduled was:	4.67	4.66	4.77
17. Lab Instructor's promptness in returning laboratory reports was:	4.19	4.30	4.36
18. The lab sessions overall were:	3.72	3.74	3.92
19. Lab Instructor's effectiveness in teaching the subject matter was:	3.96	4.02	4.16
20. Lab Instructor's contribution overall to the course was:	4.08	4.13	4.28
Course level	1.47	1.64	1.66
Challenge level	4.10	4.15	4.10
Motivation index	0.67	0.57	.59
Expected grade	3.28	3.24	3.25
	(n=74)	(n=77)	(n=78)