10-1-2003

An Analysis of Fall, 2002 Course Evaluations

Richard Frye
Western Washington University

An Analysis of Fall, 2002 Course Evaluations

Introduction

This report examines summary data from 1067 courses evaluated by the Testing Center in Fall quarter, 2002. Because teaching evaluations are voluntary, these courses represent a self-selected, non-random sample of some 48% of the 2,221 courses offered at Western during the quarter, and may be biased. Without a comprehensive norming process, the degree and direction of any bias is uncertain, and caution is advised in drawing any conclusive inferences from the averages presented here.

With this caution in mind, as shown in Table 1 for one representative item, *instructor effectiveness*, the sample data shows significant variation in evaluation ratings associated with differences in *class format, students' reasons for taking the course, course level, motivation level, expected grade,* and *general subject area*. These findings are consistent with national research findings; similar patterns exist for *course overall* and *instructor contribution*. Each of these sources of variation is discussed in more detail below.

Table 1. Sources of variation in mean scores for "instructor effectiveness"

Overall mean	**4.03** *(median = 4.17; SD = .66; interquartile range = 3.69 to 4.5)*						
Class format	***Seminar***	***Skills***	***Small lect***	***Prob solv***	***Lab***	***Lrg lect/hw***	***Lrg lect***
Fall 2000	4.28	4.12	4.05	3.98	3.96	3.85	3.78
Winter 2001	4.32	4.14	4.04	3.98	3.98	4.04	3.88
Fall 2002	4.17	4.21	4.06	3.95	4.02	4.13	3.86
Class reason	***Elective***	***Major***	***Minor***	***GUR***			
Fall 2000	4.26	4.10	4.08	3.79			
Fall 2002	*4.44*	*4.18*	*4.37*	*3.98*			
Course level	***100***	***200***	***300***	***400***	***500***		
Fall 2000	3.90	4.06	4.00	4.10	4.25		
Fall 2002	4.04	4.15	4.14	4.27	4.38		
Motivation level	***Very low***	***Low***	***Moderate***	***High***			
Fall 2000	3.82	3.77	3.92	4.18			
Fall 2002	3.69	3.85	3.97	4.22			
Expected grade	***<2.1***	***2.1-2.4***	***2.4-2.6***	***2.6-3.0***	***3.0-3.3***	***3.3-3.6***	***>3.6***
Fall 2000	2.18	2.94	3.62	3.83	3.97	4.09	4.36
Fall 2002	-	2.86	3.63	3.73	3.99	4.12	4.37
Subject area	***Psych***	***Educ***	***Humanities***	***Soc sci***	***Science***	***Engineering***	***Business***
Fall 2000	4.24	4.19	4.13	3.98	3.97	3.94	3.76
Fall 2002	4.28	3.98	4.19	4.07	3.95	4.07	3.82

Course evaluation data deviate significantly from the normal distribution. Skewness of the data, suggested by the difference between the overall mean (4.03) and median (4.17) scores, indicates that scores are substantially more clustered at the high end than would be expected in a normal distribution. High variability is evident in the standard deviation of .66, and in the interquartile range (between the 25th and 75th percentile ratings) of .81. That is, half the ratings fall between 3.69 and 4.5, with the lowest 25% below 3.69 and the highest 25% above 4.5.

Class format options

Since 1994, Western has offered seven different teaching evaluation forms, each with a different question set. Although each form was designed for a specific class format, faculty 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 2002 quarter, showing considerable variation in class size for each form. For purposes of this analysis, it is assumed that actual class formats matched the evaluation forms used.

As shown in Table 2, over a third of classes were *small lecture* classes. About a sixth were *large lecture* classes, and another sixth were *seminars*. About an eighth were *skills acquisition* classes; about a tenth were *problem solving* classes; fewer than ten percent were *labs*. Generally class size distributions for each format are skewed to the left (i.e., a longer right tail than a normal distribution). Together with large interquartile ranges (interval from 25th to 75th percentile), this suggest that within each format there are more smaller classes than larger, with substantial variability among class formats.

Table 2. Distribution of evaluations by class format

Form/ type of class	*N*	*Valid Percent*	*Mean class size*	*Std. Dev. of class size*	*Median size*	*Interquartile range*
A: Small lecture	366	35.7	29	28	25	16 -35
B: Large lecture	170	16.6	84	66	65	45 - 98
C: Seminar/ Discussion	159	15.5	20	17	17	12 -25
D: Problem Solving	92	9.0	26	15	23	17 -26
E: Skill Acquisition	132	12.9	21	24	18	9 -25
F: Large lecture/homework	28	2.7	54	31	53	37 -61
G: Lab section	78	7.6	27	13	24	22 - 27
Total	1025	100.0				

Variation in ratings by class format

While numerous questions occur on more than one evaluation form, only three questions appear on all forms. These are *course overall, instructor's teaching effectiveness,* and *instructor's overall contribution*. In addition, all the forms have a question that asks about the “challenge level” of the course, although the wording may vary slightly. Ratings on such "summary" questions have been shown in national research to be valid measurements of students' overall ratings of courses and instruction, and to be distinct from more specialized questions about course specifics.

Since these more specific questions are not common to all formats, they will not be analyzed here, but average ratings for all questions across all forms are presented in Appendix A. Summaries of average ratings on the three common questions, along with the average grade students expected in those courses, are presented in Table 3 and in Figure 1. *The average ratings are significantly different from each other both by class format..*

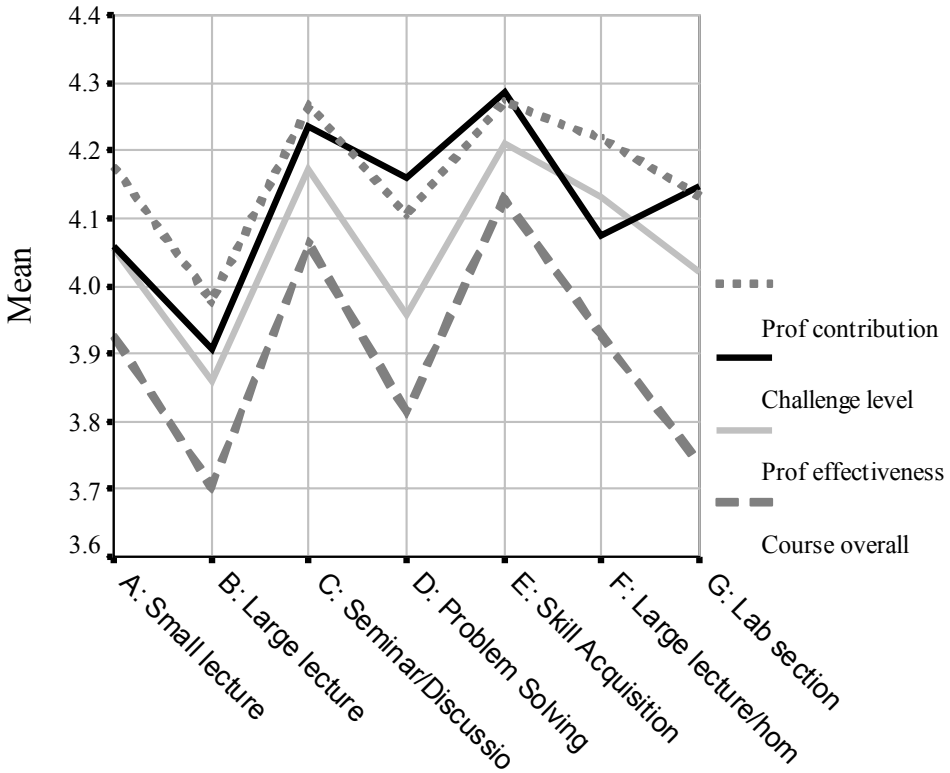
In general, seminar classes and skills acquisition classes garner the highest ratings on each summary question. Small lecture classes and large lecture classes with homework score lower and comparably to each other, while problem solving classes get lower

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. The overall mean for each question tends to approximate the "small lecture" mean because it is the most common format.

Table 3. Average ratings on common questions, by class format

Course format	*Course overall*	*Teaching effectiveness*	*Instructor overall contribution*	*Average challenge*	*Expected grade*
Over all formats	3.91	4.05	4.16	4.11	3.33
A: Small lecture	3.93	4.06	4.18	4.06	3.31
B: Large lecture	3.70	3.86	3.98	3.90	3.13
C: Seminar	4.06	4.18	4.27	4.24	3.56
D: Problem solving	3.82	3.96	4.12	4.16	3.26
E: Skills acquisition	4.13	4.21	4.27	4.29	3.57
F: Large lect/ homework	3.93	4.13	4.22	4.07	3.07
G: Lab	3.74	4.02	4.13	4.15	3.24

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



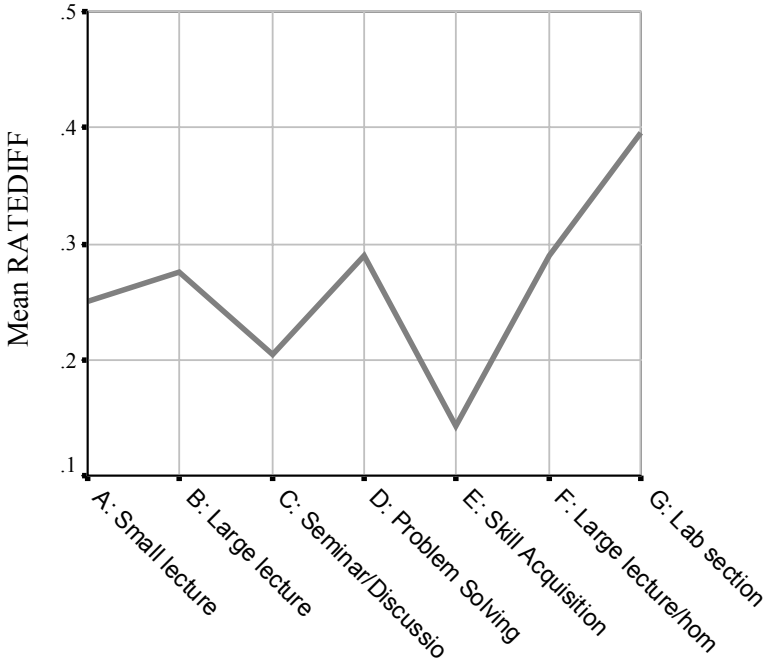
Differences of about .2 between corresponding average ratings for different class formats are statistically significant. This indicates a consistent hierarchy of average scores by class format, although adjacent formats may not be different enough to be statistically significant. For example, the average *course overall* rating for *small lecture classes* (4.06) is significantly different from *large lectures* (difference = .23) and *skills acquisition classes* (difference of -.20), but not quite significantly different from *seminar* classes (difference of -.13) or *labs* (difference = .19). A tentative hypothesis is that smaller, more interactive formats (seminars, skills acquisition) appear to earn higher ratings, other things being equal, than larger, less interactive classes (large lectures). These differences tend to persist over time, with limited variation.

Variation among questions within class formats

The nearly parallel lines in Figure 1 and correlation coefficients in the range of .94 to .97 among the ratings on the three common questions indicate that ratings on these questions are highly correlated with each other. The apparent consistency of the relative rankings of (in descending order) *instructor's contribution, instructor's effectiveness,* and *course overall* across class formats is an interesting finding, suggesting that on average students make consistent distinctions among these questions across class formats.

Although these numbers mask a great deal of individual variation, nevertheless for about two thirds (68%) of courses evaluated, average instructor ratings for both *instructor contribution* and *instructor effectiveness* are higher than ratings for *course overall*, and for 87% of courses, at least one instructor rating is greater than the course overall rating, Students seem typically more critical of courses than of instructors.

Figure 2. Rating Difference: Prof contribution - Course overall



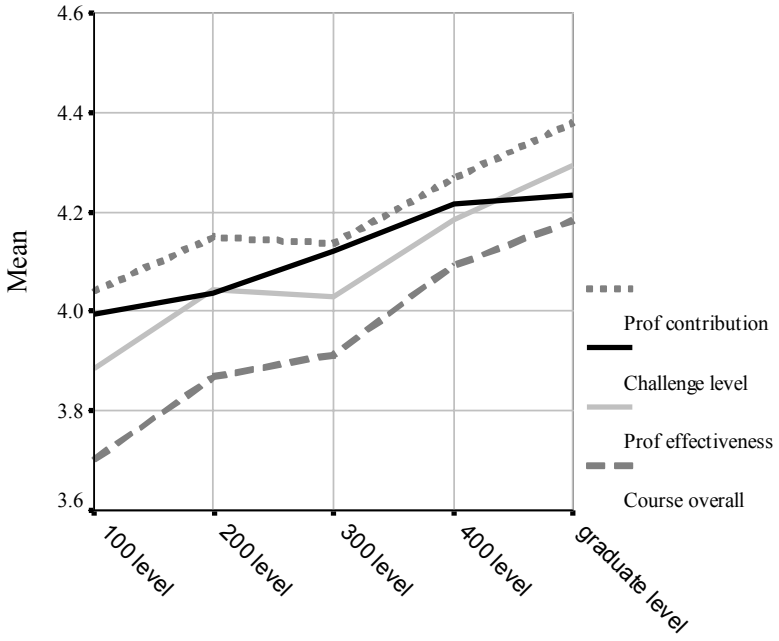
Indeed, students in less than five percent of courses rate *course overall* higher than both *instructor effectiveness* and *instructor contribution*, and only in about one fifth of classes is course overall rated higher than even one of the instructor ratings. Therefore it is a relatively rare event for the course rating to be higher than either or both instructor ratings. Such courses are more likely to be lower division (perhaps GUR) courses, and tend to have lower than average ratings on all three questions, and may indicate lower than average quality of instruction. That is, higher ratings for the course than for the instructor are unusual enough to be attention-getting; they indicate that students liked the material, but not the instruction for some reason.

While the three questions common to all evaluation forms are very highly correlated, Figure 1 shows that the differences among the three ratings do vary somewhat according to class format. This is shown in a different way in Figure 2, which shows the difference between the ratings for *instructor's overall contribution* (generally the highest of the three) and the *course overall* (generally the lowest of the three). This difference might measure the "instructor intensivity" of a format; the higher the difference in instructor and course ratings, the greater is the apparent impact of the instructor's ability over, say, course content or structure. Lab ratings, for example, appear to be relatively instructor-dependent, while skills courses seem more structure-dependent. Alternatively, this difference may simply represent the relative popularity of different formats, with skills classes and seminars being the most popular, labs being the most unpopular.

Variation by course level

As shown in Figure 3, evaluations show significant differences in ratings by course level. Students clearly give lower ratings to 100 level courses than to other class levels, and they tend to rate 200 and 300 level courses similar to each other, but significantly higher than 100 level. Senior level (400) courses get consistently higher ratings than lower division courses, and graduate courses garner the highest ratings.

Figure 3. Mean ratings by course level



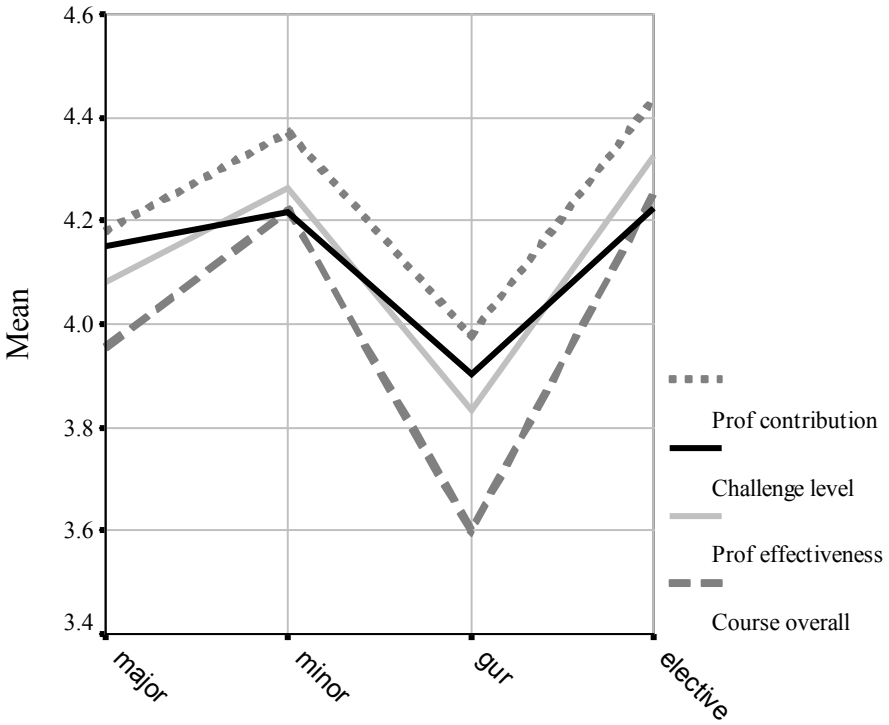
National findings confirm differences by course level, but not necessarily by class level of students. Since most classes are open to all levels of students, it is not reliably possible to isolate differences by class level. However, a rough approximation produced similar findings to those shown in Figure 3, with sophomores giving slightly lower ratings than freshmen.

Variation by reason for taking course

About two thirds (66%) of the courses evaluated were major requirements; about a fifth (20%) were GUR courses; and about 5% each were electives or minor requirements.

Ratings on the three common questions varied substantially 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, and general education requirements earned much lower ratings than courses in all other categories.

Figure 4. Mean ratings by reason for taking class



These findings are consistent with national findings, which suggest 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 (such as a Likert scale) student interest in taking a course might be useful.

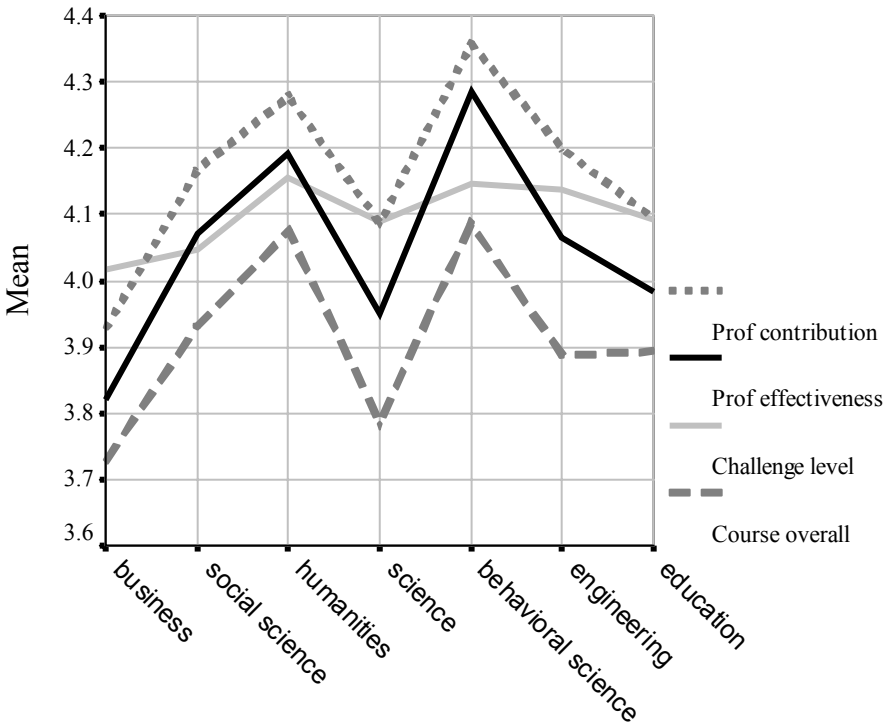
Variation by general subject area

Evaluation ratings vary significantly by subject area, as shown in Figure 5. Courses in behavioral science and the humanities get the highest ratings, followed in close order by courses in social sciences, education, physical sciences, and engineering, with business courses consistently and significantly earning the lowest ratings. These findings are very consistent with national findings, with the minor exception that at Western business courses are rated at the bottom, whereas nationally they are ranked between social and physical sciences.

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 increasing numbers of courses to complete 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; the 2002 ratings are very similar to those in 2000 except that education course ratings dropped from the higher tier (with behavioral science and humanities), to the second tier with science and engineering courses.

One partial explanation may be that different subject areas tend to favor different class formats. For example, only 7% of business courses are seminars or skills courses (the highest rated formats), compared with 8% in sciences, 27% in social sciences, around 40% in humanities, behavioral science, and engineering, and 57% in education.

Figure 5. Mean ratings by general subject area

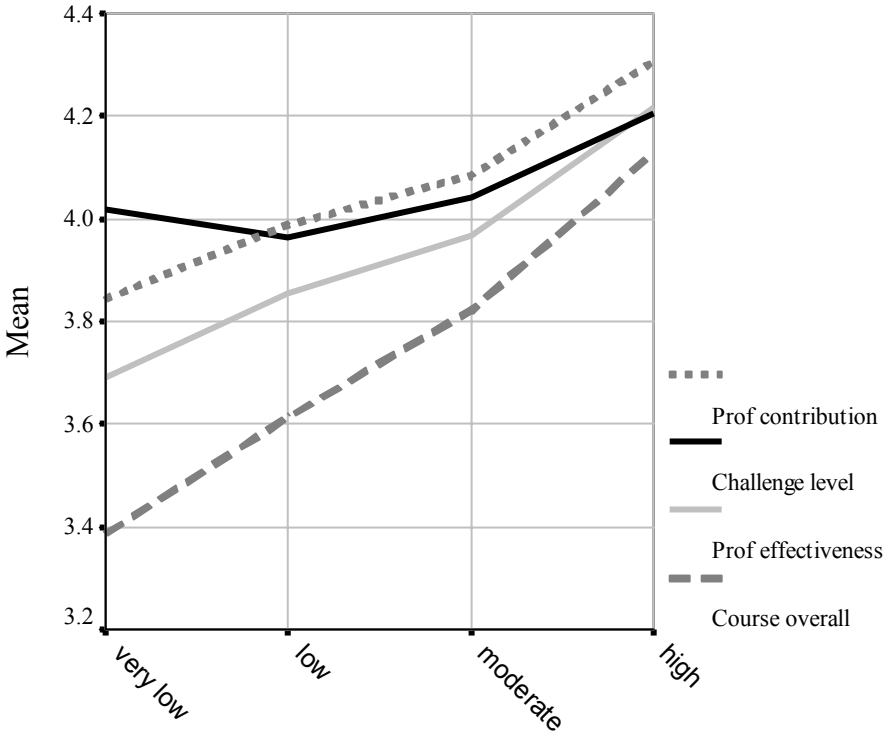


Motivation

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 yes, no, and neutral responses. 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 .71, with a standard deviation of about .25.

Ratings on the common questions show modest but significant correlations with the motivation index, between .33 (prof contribution) and .45 (course overall); the higher the index, the higher the ratings. For example, the roughly 81% of courses with indices above .5 had evaluation scores on the three questions in the range from 4.01 (course overall) to 4.22 (prof contribution), while the 19% with motivation indices of less than .5 had average scores on the same questions in the range of 3.5 to 3.92. Clearly, Western students give higher ratings to courses they are motivated to take; this is consistent with national findings.

Figure 6. Mean ratings by motivation level of students



Challenge level of course

National data suggest that, contrary to popular beliefs among faculty, students give higher ratings to more challenging courses. 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," rather than on a scale of difficulty. Across all evaluations, the average rating on these similar questions is 4.12.

The average challenge ratings are highly correlated (around .7) with the three common questions, and vary across class formats in similar fashion to the common questions, as shown in Figures 1, 3, 4, 5 and 6. Clearly, though challenge level is highly correlated with the other ratings, it exhibits a smaller variance, suggesting that the challenge levels are perceived by students to vary less across courses than course or instructor ratings.

Variation with expected grade

As was shown in Table 3 above, expected grades vary significantly by class format in the same order as overall class evaluation ratings, with rather modest correlations between expected student grades and instructor ratings, in the range of .37 to .42. Does this mean 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 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 different course formats.

Expected grades and actual grades

Because there is some controversy over the correlation between evaluation ratings of courses and instructors and expected grades, it is useful to explore the relationship between expected grade and actual grade earned. A randomly selected sample of 200 courses from Fall quarter 2000 was selected for such an exploratory analysis.

Results show 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 student average actual grade can be somewhat reliably estimated from average expected grade:

actual final grade = (1.1 x 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. And nationally, students generally benefit from the feedback provided by more frequent graded assignments, which improve learning through regular feedback about their progress toward course goals.

Deviations from normality

Distributions of responses to the common questions (including "average challenge level," which appears in various guises on all forms) are non-normally distributed. All are skewed to the left, implying that higher marks tend to "pile up" on the right end of the distribution, while lower ratings tend to be more spread out at the lower end; and all display significant amounts of negative kurtosis, implying that observations are relatively highly clustered, with larger percentages under the tails than in a normal distribution.

C*ourse overall* responses are the most normally distributed of the questions common to all evaluation forms, and show the least skewness and kurtosis. *Average challenge level* is even less skewed, but exhibits more kurtosis, with values more concentrated around the central mean. The two questions relating more directly to instructor performance, *instructor effectiveness* and *instructor contribution,* both exhibit substantial skewness (lower ratings more widely distributed than higher ratings) and kurtosis (all ratings more tightly distributed than a normal distribution).

These differences vary somewhat by class format. Most formats are about equally skewed, except for labs and large lectures with homework, which are nearly normally distributed. Similarly, most formats are also about equally concentrated about the mean, except for seminars, which are more concentrated, and labs and large lectures with homework, which are relatively normal.

The implication of these somewhat esoteric elements is that in general, evaluation ratings are more clustered toward the mean than a normal distribution, and have a compressed "right tail." Unless evaluations are generally inflated, the implication is that students regard the overall quality of instruction at Western to be quite good.

Form Discrimination

It would be useful to know how well the questions on the different forms discriminate among different qualities of course and instruction. Factor analysis of the question sets on the seven forms suggested that most of the separate items on all forms seem to measure essentially one inter-related instructor-course composite, while remaining items cluster around a few identifiable sub-elements.

Both the *small lecture* and *large lecture* forms (A, B, F) seem the least discriminatory, with nearly all questions clustering together as similar measurements of a single factor, distinct only from *challenge level* and *meeting as scheduled*. The *seminar* form seems to do a slightly better job at differentiating two separable factors; the first includes most of the questions, and the second includes *encourages self-expression, fairness of evaluation procedures, openness to student views, and support for partnership in learning*, all questions which generally do not appear on other forms, and which seem to say something important about *fostering conditions for learning*.

Questions on Form D, problem solving class, cluster into the three slightly more balanced groups, with a second factor primarily about issues related to *course organization*; and a third about the instructor's *administrative practices*. This result may be primarily an artifact of the type of class (or student in such classes) rather than evidence of differences associated with the questions asked. Form E, for skills acquisition classes, offers little discriminatory power, but does distinguish a separate element associated with *tailoring instruction to varying skill levels.*

Form G, labs, discriminates fairly well among several different instructional factors. The *lab overall* rating seems correlated primarily with *course organization and implementation.* The *instructor effectiveness* ratings seem primarily correlated with *what the instructor did and how*. A third factor unites elements associated with *clarity and fairness issues*; and the fourth factor is about *punctuality and availability*.

Overall, based on the results of factor analysis and correlation analysis, none of the forms is structured particularly well to promote discrimination among different elements of instructional quality, although the lab form, skills form, and seminar form appear to elicit distinctions not available on the other forms. Future questionnaires might include some of these elements. In addition, the research literature offers several suggestions for constructing more valid evaluation instruments:

a. Cluster items according to learning objectives or key instructional components;
b. Include questions about both content and process
c. Include items in which students enumerate actual positive or negative behaviors (theirs or instructor's) as well as opinions or satisfaction;
d. Link comments to specific questions or clusters;
e. Make questions less ambiguous;
f. Use first person wording to elicit personal responses.

Summary

Analysis of course evaluations from several quarters suggests that significant variation in course and instructor ratings are associated with many factors *other* than the quality of teaching of individual instructors, and these sources of variation are persistent over time. When using evaluations to make inferences about teaching ability or course quality, it is essential to consider the specific context of each course.

As shown in Table 1 above, evaluation ratings are sensitive to 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 substantially lower ratings to occur in a large, required GUR lecture course of a technical nature in business, engineering, or science, than in a small, senior or graduate elective seminar in humanities or psychology.

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.

One fairly clear impression is that students seem to 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 as general "best practices" for improving 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, even though this information is confounded by interaction with other variables.

Given that evaluation ratings are more concentrated about the mean than normal; that distributions of 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, given that individual instructors are likely to have more affinity for some courses, levels, or formats, low evaluations might as well be an indicator of a misallocation of faculty resources among courses within a program as a measure of an overall lack of individual ability.

Finally, because of the number and magnitude of confounding factors, there is no reliable way to separate their influence from instructor ability in any particular evaluation. Therefore, it is strongly recommended that use of teaching evaluations as a basis for making inferences about individual teaching effectiveness be heavily supplemented with additional information such as peer evaluations, course portfolios, teaching portfolios, or other independent instruments.