

Course Evaluations: An Overview

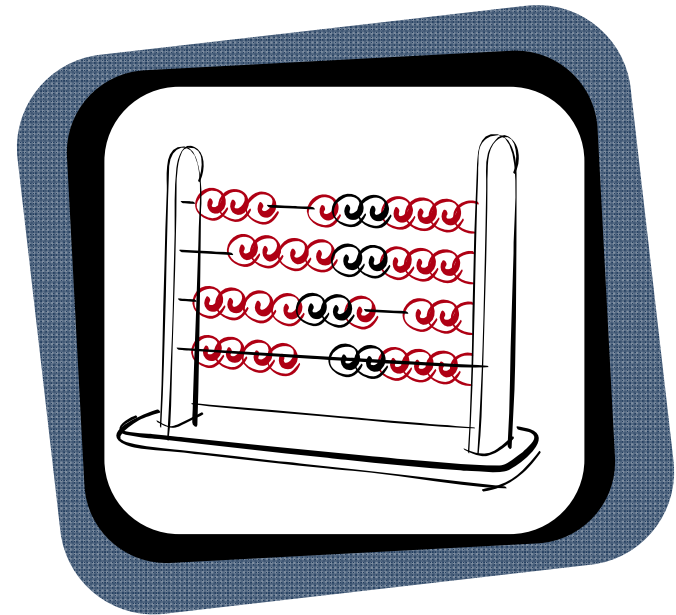
Contact Us

Michael Stob
Mathematics & Statistics Dept.
North Hall
1740 Knollcrest Circle SE
Grand Rapids, MI 49546-4403

Phone: (616) 526.7114

Fax: (616) 526.6501

Email: stob@calvin.edu



Calvin College

Summer 2008

*A brief overview of the process used to evaluate courses at
Calvin College and how to interpret the results.*

Table of Contents

HISTORY	3
THE IOTA SYSTEM	4
THE SPRING 2008 ADMINISTRATION	4
RESPONSE RATE	4
GLOBAL QUESTIONS	5
SUMMARY STATISTICS BY SECTION	6
ALL-COLLEGE NORMS.....	6
VARIATION	7
FACTORS THAT AFFECT STUDENT EVALUATIONS	9
ROLE OF COURSE IN ACADEMIC PROGRAM	9
SIZE OF CLASS	10
LEVEL OF CLASS	10
GRADES	10

following table gives the average of the ratings on Q17, given by all students for each expected grade.

Exp Grade	Ave Rating	Exp Grade	Ave Rating
A	4.24	C	3.43
A-	4.13	C-	3.41
B+	4.05	D+	3.65
B	3.92	D	3.64
B-	3.76	D-	2.56
C+	3.58	F	2.83

Since fewer than 1% of all students expect to receive a grade lower than C- (93% of the responses indicate that the student expects a B- or better in the class being surveyed), the last four entries in this table should not be scrutinized too closely. But except for these entries, it is clear that the lower the expected grade, the lower the evaluation, on average. It is interesting to note that the mean expected grade of the 10,070 students responding was 3.39.

Of course the above table, while injecting a note of caution into interpreting student ratings without understanding the context of grades, does not necessarily prove that students give low evaluations because they expect lower grades. One might just as easily interpret the results to say that the students who are being taught the best (and so are learning the material well enough to get good grades), realize that they are being taught well and answer accordingly. One might also note that the grades are distributed quite differently in different courses and, especially, in different departments. Thus, it is impossible to isolate the effect of expected grades alone in interpreting these results. One should also note that the differences in expected grades from section to section are rarely so great that this factor explains much of the variation between sections.

being taken as an elective—almost every course counts for something in a student’s program. (This is exactly the percentage of elective-taking as noted in the 2003 IAS survey.)

Size of Class

Contrary to most hypotheses, the size of the class seems to have only a small effect on overall evaluations. That is, on average, instructors of large classes receive similar evaluations to those of small- or medium-sized classes. Indeed, the mean evaluation of instructors in classes of size greater than 30 is 4.05 (Q17), slightly above the all-college average. Instructors of classes of size less than 10 do receive slightly higher ratings, on average (4.18 on question 17).

Level of Class

An interesting fact is that students in higher-level courses do not tend to rate instructors more highly than those in lower-level courses. This is quite a change from the IAS form on which students tended to rate 300-level courses considerably higher than others. The following table reports the mean rating (by section) on question 17.

Level	Mean rating (Q17)
100-level	4.01
200-level	4.07
300-level	4.03

Grades

No hypothesis is more commonly held than the one that supposes that students who receive poor grades give lower evaluations. On all IOTA forms, students report expected grades. So it is possible to correlate expected grades with individual student responses. The

History

Prior to the spring of 2008, the college used a course evaluation tool, the IAS (Instructional Assessment System), provided by the University of Washington. Students filled out a bubble-sheet of 32 questions and also answered four open-ended questions on a separate sheet. These forms were filled out during a class period near the end of the semester. While many were satisfied with the IAS system, several problems surfaced in the four years of using it that limited its usefulness. Concerns with the IAS system centered on three major issues:

1. the responses to open-ended questions were separated from the bubble-sheets in processing so that faculty members could not associate low or high scores with explanatory comments.
2. the forms were too long, frustrating students and taking too much class time.
3. the limitations of the pencil-and-paper system meant that not all courses could be evaluated each semester and that the turn-around time for numerical reports was often six weeks or more.

During the 2006-2007 academic year, the Assessment Committee studied the issue of course evaluation and recommended to Faculty Senate that a web-based system provided by IOTA be adopted. At the same time, the Assessment Committee recommended that all course sections be evaluated each semester. Faculty Senate approved these recommendations in December of 2008.

The IOTA System

The IOTA “form” consists of four open-ended questions followed by 13 questions about the course (with five possible responses each). Five questions about respondent characteristics complete the form. Additional questions may be added on a section-by-section basis. Students are invited by email to complete the form during the last full week of classes. The evaluation period closes when the final exam period begins.

Faculty members have access to all evaluations after they submit their final grades for the semester. Faculty members may see the individual responses and also summary statistics for each of their classes. Department chairs and deans have access to the evaluations for all courses in their domain of supervision.

The Spring 2008 Administration

The IOTA system was used on a trial basis during the fall semester of 2007, and it became the official system during spring 2008. All sections of all courses were to be evaluated during spring 2008. There were 17,862 possible student responses (from 3,885 different students). There were 10,070 actual responses for a response rate of 56.4%. The data in the remainder of this report refers to this administration.

Response Rate

The Assessment Committee had hoped for a response rate of at least 70%. (The IAS system regularly resulted in response rates of 80-85% as it was administered to a captive audience.) Of the 865 sections evaluated, 194 reached the 70% benchmark. There were 261 sections with response rates of less than 50%. It is important to note that there was no institutional reward for responding nor were there any sanctions for not responding. Faculty Senate had

the instructor (question 17 is about the instructor, after all). However, the table also clearly indicates that some of the variation is due to the particular course being taught, and much of the variation isn't really attributable to the department, course, or instructor. Note that 25% of the time an instructor's scores in two different sections of the same course differ by more than .4. In other words, an instructor of Religion 121 might receive a rating of 4.0 from one class and 3.6 from another, and we should not be particularly surprised by that.

The above analysis clearly has consequences for the use of IOTA data in supporting personnel decisions. We should be very careful in making any firm conclusions about student evaluation of instructor effectiveness from just one or a few classes. We should collect data from many different sections and from several different courses, if possible.

Factors That Affect Student Evaluations

It is easy to make guesses about what factors are liable to affect the ratings that one might get in a particular class. For example, it is generally argued that students in a 100-level core course will on average tend to rate instructors lower than students in a 300-level course for majors.

Role of Course in Academic Program

Students who take a course as a major requirement, a minor requirement, or a core requirement do not give substantially different ratings overall. That is, the distribution of responses on the global questions for, say, the totality of students taking courses for core, is very similar to that of students taking a course for a major. On the other hand, students do tend to give higher ratings to instructors in courses taken as electives and lower ratings in courses taken as a required cognate for a major. One finding in this regard is that only 6% of the responses indicate that the course is

factors such as the time of day or the temperature of the classroom. Of course we would like to know, if the ratings of two instructors differ, how much of that difference to attribute to the instructor. This is not a question that is possible to answer without running experiments to control for all other effects. However, we can get an estimate of the proportion of variation attributable to the instructor by observing what happens in a variety of situations, such as the cases when the same instructor teaches two different sections of the same course. The following table gives some information about the amount by which the mean score on question 17 (rating of instructor) differs between two classes of a certain sort. (Again, only sections with at least 10 respondents were considered in this analysis.)

Differences in means between two sections (Q17)

	25%	50%	75%
different courses and instructors	0.92	0.52	0.24
same department	0.88	0.49	0.23
same course	0.85	0.48	0.20
same instructor	0.48	0.30	0.11
same instructor and course	0.40	0.21	0.10

The entry .52 in the first row means that if one were to choose two sections of different courses with different instructors, 50% of the time the means of those two sections on question 3 would differ by more than .52. On the other hand, if one knew that the two different sections were of the same course and taught by the same instructor, 50% of the time the difference would be less than .21.

What this table tells us is that much of the variation in responses, by different sections of students, to question 17 is attributable to

adopted a recommendation to withhold grades from students who did not respond, but that recommendation proved too difficult to implement. Further work needs to be done to identify strategies for increasing the response rate.

Global Questions

In analyzing the overall pattern of student responses, we focus here especially on questions 9 and 17. These questions are the global questions that focus on the overall evaluation, by the student, of the course and instructor. (The literature on the use of course evaluation forms consistently suggests that only global questions should receive much weight in personnel evaluation settings.)

Question 9	The course as a whole was
Question 17	The instructor was

Each question had the same choices of response: Excellent, Very Good, Good, Fair, Poor. Students tended to rate both the course and instructor highly and, overall, to rate the instructor higher than the course. The following table gives the frequency of each response on these two questions.

	E	VG	G	F	P
Course as a whole	25.7%	35.7%	26.6%	9.7%	2.4%
Instructor	41.2%	31.7%	18.5%	6.7%	1.9%

The IAS system had four global questions and a six-response scale (adding Very Poor). The response percentages for each of the global questions for the 2003 calendar-year administration of the IAS form (all sections of all courses) is here for comparison. It's clear that the pattern of responses is quite similar for the IOTA form, and the differences could be attributable to the availability of the extra choice on the IAS form.

	E	VG	G	F	P	VP
Course as a whole	24.0%	37.4%	28.0%	8.3%	1.9%	0.4%
Course content	21.8%	38.6%	30.7%	7.4%	1.3%	0.2%
Instructor's contribution	38.6%	34.1%	19.3%	6.1%	1.5%	0.5%
Instructor's effectiveness	33.4%	33.2%	21.5%	8.7%	2.5%	0.7%

Summary Statistics by Section

Each instructor is provided with a report for each of the sections that he or she taught. For each question, the instructor receives both the percentages of students using each response and also the mean response for each question. The mean is computed by converting the five responses to the numbers 5, 4, 3, 2, 1 and computing the numerical average. The value 5 is assigned to the most favorable response. Since the distribution of responses is heavily skewed (toward 5), the mean can be heavily influenced by a few negative responses. Therefore, in small sections, a response of Poor has great influence on the mean, and this should be taken into consideration when comparing means across sections, particularly small sections.

All-College Norms

The following table summarizes the distribution on questions 9-17 of the 523 sections for which there were 10 or more students responding. The table gives the percentiles of the mean score by section on each question. In other words, the first entry in the table

2.91 indicates that in 10% of such sections (about 52), the mean score on question 5 was 2.91 or lower. Similarly the third entry, 3.73, indicates that in half the sections, the mean score on question 5 was 3.73 or lower.

	10%	25%	50%	75%	90%
Q9—Course as a whole	2.91	3.39	3.73	4.08	4.38
Q10—Teaching methods	2.86	3.33	3.81	4.20	4.50
Q11—Enthusiasm	3.55	4.00	4.33	4.59	4.79
Q12—Organization	2.90	3.51	3.93	4.27	4.50
Q13—Clarity	2.92	3.48	3.87	4.19	4.47
Q14—Helpfulness	3.20	3.67	4.00	4.30	4.60
Q15—Fairness	3.21	3.60	3.93	4.23	4.42
Q16—Prompt feedback	3.00	3.50	3.87	4.20	4.47
Q17—Instructor	3.20	3.69	4.10	4.45	4.67

It is clear that these numbers indicate that Calvin students rate their courses and instructors very highly.

Variation

In using data from IOTA summaries, it is important to understand that not all differences between sections are significant or necessarily indicative of a difference attributable to the instructor. The means of two different sections may differ due to the instructor, the particular course being taught, the particular group of students that took the course (we've all had particularly "bright" classes and perhaps some of the other kind), or a myriad of other