Grinnell College
Institutional Research

Project Report
October 2000

Carol Trosset, Director
Scott Baumler, Senior Research Analyst

End-of-Course Student Ratings of Instruction

For several years at Grinnell each academic department used its own end-of-course evaluation form. Many departments had more than one form for use with different courses. The Office of Institutional Research (IR) produced, distributed, and scanned the forms. Summary statistics and copies of the forms were sent to the instructor and to the instructor's department chair, who reported on overall performance in salary recommendation letters.

The wide variety of questionnaires led to difficulties for the Budget Committee and the Personnel Committee in dealing with numeric information from end-of-course forms. Laura Sinnett, acting for the Executive Council, carried out an investigation of the end-of-course forms. In March 1999 Sinnett presented a report analyzing student ratings of teaching. She utilized data from faculty members who agreed to share their spring 1998 course forms. Sinnett identified questions that were asked in a similar fashion across several departments and performed statistical analyses to test the reliability and validity of the instruments. Though she did not identify any major biases in the forms used, she cautioned the faculty about the limited comparability of the various forms (and, therefore, the limited legitimacy of using those scores for performance and salary reviews).

IR determined that producing and processing the 42 different forms in use in Fall 1998 took one-third of a full-time staff member. The process also required a great many other resources which did not appear justified by the low quality of the data then being obtained. The Dean of the College also felt that more efficient use should be made of the IR staff's work time.

Based on these findings, the Executive Council recommended, and the faculty adopted, a one-semester test of a single college-wide form. The form had six questions and a six-point Likert scale. Other questions gathered information about the student's gender, major, class year, and expected grade. Comments were also invited on each question.

The form was pilot-tested in one of Sinnett's courses prior to the end of the spring 1999 semester. This was done as a check on the physical design of the survey instrument and to ensure that the sequence of response categories did not systematically affect the results obtained. The form was then mass produced and used campus-wide as the standard instrument. Instructors were encouraged to augment the standard form with their own questionnaires as they felt necessary.

IR performed both quantitative and qualitative analyses of the results. Their document entitled *End-of-Course Evaluations: Spring 1999 Results* was presented in October 1999, and is available on the office's web page. As a result, the Executive Council made slight revisions in the form and recommended a year-long follow-up test, approved by the faculty on November 1, 1999.

The rest of this report summarizes the findings of the follow-up study.

Overview

Student ratings from 610 individual course-sections were collected during the 1999-2000 academic year. The yield rate for the student questionnaires (number of completed forms ÷ total enrollments) was 89 percent. The response rate for the instructor surveys was 81 percent. Of the instructors who responded, 13 percent indicated that they had used their own end-of-course questionnaire in addition to the standard form.

Five-hundred of the 610 courses garnered at least a 75 percent share of students who indicated they *moderately* or *strongly agreed* that they learned a lot in the course. Across all students, all courses, and all questions, 82 percent of the valid responses were *moderately* or *strongly agree*. Table 1 provides a summary of the student responses.

The positive responses seem to indicate that, in the aggregate, students generally think highly of instruction at Grinnell. These findings are corroborated by the results of our senior surveys. Every year, 85 to 98 percent of graduating students indicate they were satisfied or very satisfied with the instruction they received here. This confirms the tendency of students to give high ratings on end-of-course forms. In other words, we have additional evidence that supports the course ratings: students are generally very satisfied with the instruction they receive at Grinnell.

Student Responses

The first five items on the end-of-course form asked students to reflect on course characteristics. The sixth item asked students for their impressions of how much they learned. With a six-point disagree/agree Likert scale and space provided for text comments, the specific items were:

Q1: The course sessions were conducted in a manner that helped me to understand the subject matter of the course.

Q2: The instructor helped me to understand the subject matter of the course.

Q3: Work completed with and/or discussions with other students in this course helped me to understand the subject matter of the course.

Q4: The oral and written work, tests, and/or other assignments helped me to understand the subject matter of the course.

Q5: Required readings or other course materials helped me to understand the subject matter of the course.

Q6: I learned a lot in this course.

It is important to remember that students can only report on what they are able to perceive in the last week of class. They can and do reflect on things that are not specifically asked about on the forms, as well as responding to the questions. However, a number of studies done elsewhere have found that student ratings are moderately correlated with independent measures of student achievement and learning (such as student exam scores, instructor self-ratings, and peer reviews of teaching).

Applying the Results

There are two basic ways to use student ratings of teaching:

1. *Formative* – provides feedback for personal learning, development, and self-improvement

2. *Summative* – the results are used, with an appropriate weighting, for personnel matters such as retention, promotion, and salary adjustments

Given that student ratings provide relevant information and consensus is reached regarding the intended functions of the evaluations, consideration can then be given to how the information is referenced. There are three basic referencing techniques:

1. *Self referencing* **–** an instructor's current performance is compared with his or her own previous performance, or other self-identified criteria.

   Formative evaluation is generally a positively-embraced concept with the idea of personal growth and improvement to benefit the instructor, the students, and the college as a whole. If instructor development is the sole purpose of the end-of-course questionnaires, locally-produced, individualized forms may be sufficient. Departmental and college-wide support is available for using student feedback to improve course design and delivery.

2. *Norm referencing* – an instructor's performance is assessed in relation to the performance of other faculty members.

   Without reference points to describe relative standing, one may encounter a "Lake Woebegone effect": a situation where everyone is above average. It is wonderful situation, if true, but the proclamation depends on what one sets as the baseline; that is, what one considers "average." While it is true that our faculty and the instruction they provide is par excellence, it may not be true that all instructors at Grinnell are equally skilled in their abilities to generate student fulfillment and contentment with respect to coursework. Ipsative or self-referenced assessments lack any external frame of reference and, as such, are devoid of meaning for people who need to make decisions based on relative standing. Norm referencing provides landmarks.

3. *Criterion referencing* – an instructor's performance (or even the college as a whole) is measured against pre-determined standards.

   For example, "We expect a minimum of 90% of the students to agree that the course sessions were conducted in a manner that helped them understand the subject matter." This is mastery approach, with room allowed for unusual or idiosyncratic circumstances. Developmental assistance and other corrective actions may be initiated in situations where the threshold is repeatedly crossed.

These possibilities, which are not mutually exclusive, should be considered when deciding how the data should be used within a holistic faculty review process. Ultimately, the approach decided upon will dictate what data to gather, how to summarize it, and how it should be disseminated.

Quantitative Aspects

Statistical summarization of the scaled *disagree/agree* items on the forms provides a standardized, shorthand method of dealing with large amounts of information. There are two common approaches with this type of data:

1. **Percentages** – the proportional frequency of the student responses, such as, "85 percent of the students *strongly agreed*." Percentages preserve the specific category marked by each student. Each response category can be reported individually or some categories can be lumped together. Potential drawbacks: a) If some of the response categories are not combined, there are multiple percentages to contend with for each question – which may be cumbersome for comparisons. b) If some response categories are combined, the scale is artificially divided. This can also result in a loss of data because the intensities of some of the responses are diluted. c) Percentages are extremely unstable for small classes.

2. **Average scores** – the response categories are coded with numerical values and the mean is computed. In the context of obtaining a composite measure from a panel of raters (students), averages help to "net out" the effects of raters who may be either lenient or severe (or otherwise biased), much like Olympic scoring. It is a way of getting a feel for the central tendency of the group.

   A downside of using averages is that the precision of the result can be over-interpreted. To mitigate this hazard, statistical margins of error (confidence intervals) can be used to communicate the *quality* of the data. In effect, confidence intervals can appropriately "dim" these scores (the means) to avoid over-emphasizing small numerical differences. In small classes the confidence intervals may span the entire range of scores. This may seem strange but it acts as a "quality assurance" tool.

   A statistic used to assess the homogeneity of course means, the intraclass correlation coefficient (ICC), indicates that these scores are quite stable. An ICC close to 1.0 indicates a very consistent measure. (See *End of Course Evaluations: Spring 1999 Results* for a complete description of this approach.) The table below provides the results for Question #6, "I learned a lot in this course."

Approach	Intraclass Correlation Coefficient	Mean-Absolute Deviation	Number of Observations
Same Course, Different Instructor	.50	.26 (5.2%)	46
Different Course, Same Instructor	.58	.25 (4.8%)	131
Same Course, Same Instructor	.82	.15 (3.0%)	74

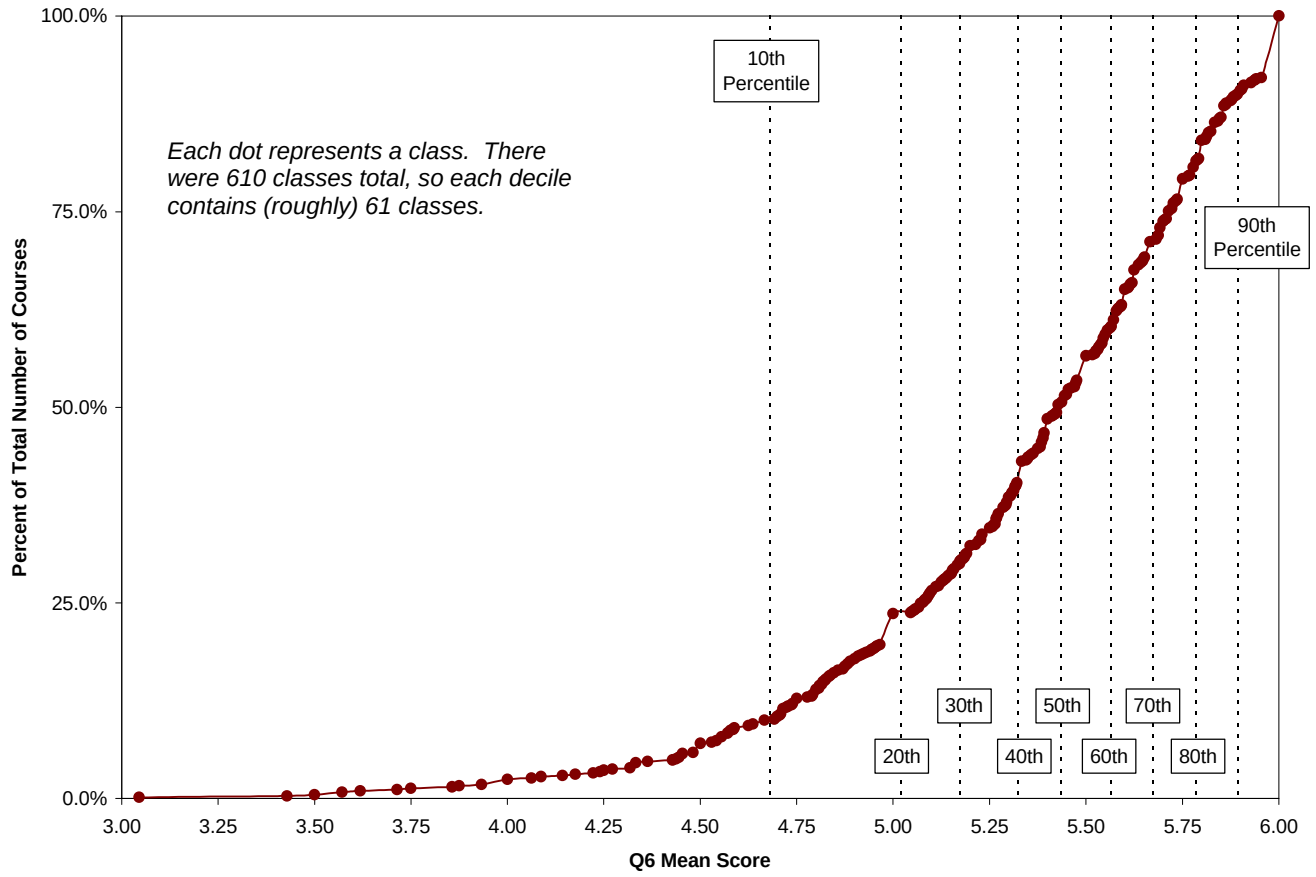
   These statistics indicate the "tightness" of the mean course scores. For cases where the same instructor taught the same course, the mean scores were typically within .15 points of each other. As in the previous study, *instructor effects appear to outweigh course effects*. The actual differences among the means, however, are relatively small.

Percentiles

Percentiles are convenient guideposts used to indicate an individual's relative position in a group. These "markers" are expressed in terms of the percentage of cases that fall below a certain score. For example, the 50th percentile for Question #6 was 5.43 – half of the course mean scores fell below this point (and, of course, the other half were above this point).

Figure 1 shows the cumulative distribution of mean scores for Question #6. Deciles are indicated with dashed vertical lines. Note how they are closer together in the middle and upper end of the scale and farther apart at the lower end. This type of "distortion" is typical for distributional data. Any particular percentage of cases in the middle of a (roughly) normal distribution will cover a shorter distance on the x-axis (raw scores) than the same percentage will cover near the tails of the distribution. One might think that the difference between a mean score of 3.50 and 4.50 would signify a significant difference – and perhaps it does, depending on the context in which the number is used – but in terms of within-groups norms both scores are still below the 10th percentile. For matters concerning relative standing, raw point scores are vacuous without normative references.

Figure 1: Cumulative Distribution of Q6 Mean Scores



Validity

Some degree of validity is confirmed by the fact that no significant biases were found using demographic characteristics of students or faculty. In other words, the form is not grossly biased for or against any identifiable class of individuals.

Our analyses involve very large quantities of data. With large samples, very small differences become statistically significant. "Statistically significant," however, does not necessarily indicate "meaningful." As such, the focus should be on effect sizes, which were generally small. Table 2 provides information from several exploratory tests.

Small tendencies at the student level (such as the tendency of first-year students to give lower marks) do tend to accrue at the course level as the relative proportion of those students in a class increases, but the effects are still modest. Subtle point differences should not be troublesome if the practice of reporting average scores in terms of confidence intervals is continued.

The Spring 1999 qualitative analysis raised some challenges to the validity of the data. Although the questions on the forms ask students to focus specifically on things that contributed to their learning, most student comments concerned other issues. In other words, students evaluate faculty on many different criteria, not just those asked for on the forms. Therefore, it is important to remember that scores do not necessarily reflect student views of the issues faculty members want measured. In fact, both extremely high and extremely low scores were both characterized by frequent comments relating to instructor personality and/or student expectations, rather than to anything about what the students got out of the course.

Validity of the instrument cannot really be assessed without some independent measure of teaching quality that does not depend on surveying students. Additionally, the question of validity depends on the use to which the data will be put. In this case, it is clearly valid to use course ratings as measures of student satisfaction. However, it is not valid to use them as direct measures of teaching quality (unless the faculty determines that satisfied students are the primary measure of good teaching).

Caveats & Observations

- Student perceptions of their own experiences are relevant to evaluation and review. Although it is safest to interpret the ratings as measures of student satisfaction rather than of teaching quality, repeated reports of student dissatisfaction would certainly be cause for concern

- It is not legitimate to make any fine distinctions between individuals. The limited quality of the data does not justify such comparisons beyond the use of confidence intervals or some other generalized classification system.

- Course mean scores are relatively stable; stable, in fact, to the point where the instruments may not be sensitive enough to pick up meaningful change over time for established instructors in established courses. Formative-related items (such as questions related to books or computers) may elicit noticeable differences, but the "overall rating" may well remain constant. New instructors may encounter "jumps" in their ratings as they gain classroom experience and hone their instructional resources.

- If the forms and results are to be used as a component of faculty performance evaluations:

  - There should be a consistent college-wide form used in all classes.
  - Some kind of benchmark data should be provided to help people interpret the scores.
  - So as not to overburden students, the forms should be short to allow instructors to use their own forms for formative purposes.

- If the forms are not to be used for performance evaluation, but rather for professional development, then a variety of comments-only forms would probably be more useful.

In deciding what sort(s) of forms to use, the resources of the Office of Institutional Research should be considered. With some outsourcing, a single college-wide scannable form can be handled by the office. However, individualized departmental forms require so much personnel time that they cannot be handled by the office under current staffing conditions.

A great deal of research has been done all over the country attempting to design and assess this sort of form. Results of these studies are contradictory, and no one has found a perfect solution. Further research at Grinnell is unlikely to produce any findings we have not already obtained in the past year and a half.

Student insights can make important contributions to both formative and summative evaluations of teaching. Student ratings of instruction are a convenient, reasonable method of communicating summary data about student perceptions. Thinking of them as "ratings" rather than as "evaluations" serves as a reminder that these data need to be interpreted. "Ratings" are data provided by students, while "evaluation" is a more comprehensive process performed by faculty peers, using a variety of approaches.* The role student ratings should play in the wider evaluative process is an issue for the faculty to decide. ❖

* Cashin, William E., *Student Ratings of Teaching: The Research Revisited*. September 1995. Idea Paper No. 32, Center for Faculty Evaluation and Development, Division of Continuing Education, Kansas State University.

Table 1: Summary of Student Responses

Q1: The course sessions were conducted in a manner that helped me to understand the subject matter of the course.

	Frequency	Percent
Strongly disagree	105	1.2
Moderately disagree	213	2.3
Slightly disagree	246	2.7
Slightly agree	901	9.9
Moderately agree	3,207	35.3
Strongly agree	4,378	48.2
Not applicable/don't know	16	0.2
No response	9	0.1
	9,075	100.0

Q2: The instructor helped me to understand the subject matter of the course.

	Frequency	Percent
Strongly disagree	90	1.0
Moderately disagree	145	1.6
Slightly disagree	196	2.2
Slightly agree	778	8.6
Moderately agree	2,663	29.3
Strongly agree	5,159	56.8
Not applicable/don't know	29	0.3
No response	15	0.2
	9,075	100.0

Q3: Work completed with and/or discussions with other students in this course helped me to understand the subject matter of the course.

	Frequency	Percent
Strongly disagree	139	1.5
Moderately disagree	215	2.4
Slightly disagree	305	3.4
Slightly agree	1,383	15.2
Moderately agree	2,817	31.0
Strongly agree	3,557	39.2
Not applicable/don't know	623	6.9
No response	36	0.4
	9,075	100.0

Q4: The oral and written work, tests, and/or other assignments helped me to understand the subject matter of the course.

	Frequency	Percent
Strongly disagree	120	1.3
Moderately disagree	183	2.0
Slightly disagree	321	3.5
Slightly agree	1,173	12.9
Moderately agree	3,170	34.9
Strongly agree	3,717	41.0
Not applicable/don't know	352	3.9
No response	39	0.4
	9,075	100.0

Q5: Required readings or other course materials helped me to understand the subject matter of the course.

	Frequency	Percent
Strongly disagree	130	1.4
Moderately disagree	215	2.4
Slightly disagree	308	3.4
Slightly agree	1,109	12.2
Moderately agree	2,893	31.9
Strongly agree	3,728	41.1
Not applicable/don't know	654	7.2
No response	38	0.4
	9,075	100.0

Q6: I learned a lot in this course.

	Frequency	Percent
Strongly disagree	122	1.3
Moderately disagree	170	1.9
Slightly disagree	205	2.3
Slightly agree	878	9.7
Moderately agree	2,645	29.1
Strongly agree	4,991	55.0
Not applicable/don't know	27	0.3
No response	37	0.4
	9,075	100.0

Table 2: Exploratory Tests for Q6

Factor (Independent Variable)	Index of Effect Size (Strength of Relationship)	Approach (Statistical Test)	N
Student-Level Analysis			
College Division	$\varepsilon^2 = .01$	Kruskal-Wallis	8,724
Student's Gender	$r_b = -.07$	Mann-Whitney	8,482
Course Level	$\varepsilon^2 = .01$	Kruskal-Wallis	9,011
Instructor's Gender	$r_b = .06$	Mann-Whitney	8,769
Times Instructor Taught Course	$r_s = .04$	Spearman	7,118
Instructor's Rank	$\varepsilon^2 = .01$	Kruskal-Wallis	7,078
Instructor's Minority Status	$r_b = .01$	Mann-Whitney	8,561
Student's Year in College	$\varepsilon^2 < .01$	Kruskal-Wallis	8,755
Course in Student's Field	$\varepsilon^2 = .01$	Kruskal-Wallis	8,887
Course Enrollment	$r_s = -.09$	Spearman	8,986
Course-Level Analysis			
College Division	$\eta^2 = .04$, $\varepsilon^2 = .04$	ANOVA, Kruskal-Wallis	591
Percent Female Students	$r = .08$, $r_s = .08$	Pearson, Spearman	602
Percent Male Students	$r = -.08$, $r_s = -.08$	Pearson, Spearman	602
Course Level	$\eta^2 = .06$, $\varepsilon^2 = .05$	ANOVA, Kruskal-Wallis	610
Instructor's Gender	$\eta^2 < .01$, $r_b = .03$	t Test, Mann-Whitney	590
Times Instructor Taught Course	$r = .04$, $r_s = .09$	Pearson, Spearman	485
Instructor's Rank	$\eta^2 = .03$, $\varepsilon^2 = .02$	ANOVA, Kruskal-Wallis	475
Instructor's Minority Status	$\eta^2 < .01$, $r_b = .06$	t Test, Mann-Whitney	576
Percent First-Year Students	$r = -.20$, $r_s = -.25$	Pearson, Spearman	606
Percent Seniors	$r = .15$, $r_s = .18$	Pearson, Spearman	606
Percent for Whom Course was in Major Field	$r = .19$, $r_s = .20$	Pearson, Spearman	608
Course Enrollment	$r = -.18$, $r_s = -.26$	Pearson, Spearman	607

where:

ε^2 (epsilon squared) ranges from 0.00 to 1.00

r_b (rank biserial correlation) ranges from -1.00 to +1.00

r_s (Spearman's rho) ranges from -1.00 to +1.00

η^2 (eta squared) ranges from 0.00 to 1.00

r (Pearson correlation) ranges from -1.00 to +1.00