

Student Evaluations and the Assessment of Teaching: What Can We Learn from the Data?

Martin D. D. Evans

and

Paul D. McNelis

Department of Economics, Georgetown University

April 2000

This study was commissioned by the University Provost's Office in Fall 1999. We wish to thank University Provost Dorothy Brown, University Registrar John Pierce, Vice President Joseph Petit, and Director of Institutional Research Michael McGuire for making available the data and for their support.

Introduction

This paper is an assessment of the limits of using data from student evaluations for the evaluation of teaching performance or “effectiveness” in undergraduate classes at Georgetown University. In particular, this study shows that the use of mean scores of “overall effectiveness” of the instructor, for judging relative performance across classes of different sizes, different levels, meeting at different times during the day, and across instructors with different “grading reputations”, is an extremely limited and even misleading way to use the survey data.

While members of different committees at departmental, school and university levels often make implicit and informal adjustments to the reported mean scores of instructors (more often by not, by simply eyeballing the data), there has been no effort to formalize how the scores should be adjusted to correct for other factors, such as class size, level, and meeting time, that are unrelated to the classroom effectiveness of the instructor. As a result, there is no evidence on how much of a given teaching performance, measured by the mean score, can be attributed to such factors, and how much of the mean score can be attributed to “teaching ability” .

Given the importance which the university places on teaching excellence in its annual merit review and in its promotion and tenure processes, this issue has major incentive effects for the willingness of faculty to teach particular types of courses rather than others. A major finding of this study is that the way we currently make use of the data from student evaluations severely distorts the way we allocate teaching resources. We also find that the current practice for identifying both effective and ineffective teaching is inconsistently applied across individuals. As a consequence, some individuals are erroneously viewed as being ineffective teachers while others are viewed as being highly effective. In short, eyeball comparisons of one instructor’s scores against another’s, are more often than not completely arbitrary.

Our findings also shed light on a possible source of grade inflation. We find that the grades given by an instructor in the past year have a sizable and significant affect on the student scores the instructor receives in courses taught this year. This “grading reputation effect” has particularly adverse consequences on the mean scores if the instructor graded severely relative to the average grade distribution across the university in the previous year. We do not investigate, or even speculate, on the origins of this reputation effect. However, our results do show that instructors who give out grades below the university average, can expect to receive significantly lower evaluation scores in subsequent semesters irrespective of how effective they are in the classroom. To the extent that individual faculty recognize this consequence of giving grades below the norm, and are concerned about the possible professional implications of low evaluation scores, there is clearly an incentive for them to think carefully before giving low grades that accurately reflect the academic performance of the students. Ignoring this “grading reputation effect”, as is our current practice, makes maintaining academic standards at Georgetown a harder task than need be the case.

The Data

The data in this study come from the University Registrar and the Director of Institutional Research. The evaluations data are for all undergraduate courses taught at Georgetown in the Spring of 1999. The information on instructor's grades comes from the 1997/8 academic year. The identity of all instructors and the names of all courses were deleted from the database before we were given access to the data in order to preserve the anonymity of instructors. The data only allow us to link a course instructor with his or her grades through an anonymous identifier that was constructed for this study.

Table I shows the cumulative distribution of the mean scores (by class) on question III. 5 of the student survey "Overall Evaluation of the Instructor". Judged by mean scores across classes, the overall teaching effectiveness of the faculty looks quite good. The average mean teaching evaluation score on overall effectiveness of the instructor is 4.3, and the median is 4.5, both well above the "official" average rating of 3.0 on the 1 to 5 scale. There are quite a few perfect and near perfect scores, while there are very few scores below 3. A score below the official average of 3 makes one an extreme outlier at Georgetown University.

Table 1: Cumulative Distribution of Mean Scores

Score Greater Than	Percent of all Scores
4.8	22.14%
4.6	41.92%
4.4	56.31%
4.2	69.12%
4	79.24%
3.8	84.03%
3.6	88.57%
3.4	90.87%
3.2	93.63%
3	95.27%

It is well known around the campus that Georgetown students receive high grades. Table 2 reports statistics on how many A/A- grades were given across all courses during the 1997/8 academic year. The average number of A/A- grades per class at Georgetown is 50 percent, the median is 48 percent. As the table shows, in less than 1 percent of all the classes taught during the year where less than 10 percent of the students awarded an A/A- grade. (There were no courses in which there are no A/A- grades.) At the other extreme, in almost 10 percent of the class at least 80 percent of the students were awarded an A or A-. Clearly, A/A- grades are common across all classes, and in some, anything below an A- is a rarity.

Table 2: Cumulative Distribution of A Grades Across Classes

Percentage of A/A- Grades in Each Class	Percentage of Classes
90.00%	4.27%
80.00%	9.53%
70.00%	18.07%
60.00%	30.49%
50.00%	46.78%
40.00%	69.91%
30.00%	87.91%
20.00%	98.75%
10.00%	99.80%

In principle, evaluation forms are filled in by all the students in each class so that the survey aggregates all views on the effectiveness of the instructor. In practice there are a large number of classes where only a subset of students actually taking the class fill in the evaluation forms. We refer to this fraction as the survey response rate. We measure the response rate as the number of evaluations actually filled in divided by the number of students enrolled in the class. Table III gives the distribution of the response rates in Spring 1999.

Table 3: Cumulative Distribution of Response Rates

Response Rate Less Than	Percent of All Classes
100.00%	100.00%
90.00%	57.56%
80.00%	27.73%
70.00%	13.27%
60.00%	5.91%
50.00%	2.10%
40.00%	1.18%
30.00%	0.26%

The mean response rate is 85 percent and the median is 88 percent. The table shows that the percentage of classes for which the response rate is less than 70 percent is 13 percent. This is a surprisingly large number when one considers that these response rates reflect attendance of students at class at the end of the term, either on the last day of class or at classes during the final weeks of the term, when the survey forms are handed out. Clearly, the surveys fall well short of aggregating the views of all students in every class.

Table 4: Cumulative Distribution of Class Size

Class Size Less Than	Percent of All Classes
5	7.23%
10	24.38%
15	43.82%
20	59.59%
25	67.67%
30	72.86%
35	77.86%
40	83.51%
45	86.33%
50	89.29%
100	99.01%

There is also considerable variation in the sizes of classes taught at Georgetown. Table 4 shows the distribution of class sizes measured by enrolment. The majority of classes are below 20 and almost 90 percent of classes are below 50. There are also a significant number of classes that are very small (i.e., less than 5 students), and very large (i.e., more than 50). The teaching environments in these classes must be very different, a fact that could affect teaching scores independently of how effective a teacher the instructor is.

What Affects Student Evaluation Scores?

We begin our analysis by asking whether factors other than the classroom performance by the instructor affect the evaluation scores. In particular, we will examine whether class size or student response rates matter. Similarly, does the grading reputation of the instructor, either as an easy or hard grader, affect the evaluation scores? As we document below, the answers to these questions is a resounding yes. Students in small classes (i.e., with 5 students or less) systematically give higher evaluations to their instructors than do students in larger classes. The scores received for classes in which all the students fill out the survey are higher than for classes in which only a fraction do. We also find that instructors who gave a higher fraction of their students low grades in the previous year will have significantly lower evaluations on all the courses they teach in the current year. A reputation for grading below the perceived university norm increases the likelihood that an instructor will be regarded as an ineffective teacher.

The remainder of this section describes the statistical methodology and presents our results in detail. The question of whether the evaluation survey contains information about the teaching effectiveness of individual instructors is addressed in the next section. Readers wishing to focus on our main results, may skip to this section without loss of continuity.

Methodology

The methodology we use in this study is widely applied in Biostatistics and medical research: It is used to determine the effectiveness of specific drugs or therapies by discriminating between “control groups”, one of which uses a new regime, and another which makes use of a placebo. Before any decisions are made about the effectiveness of the drug, the results from both groups are “adjusted” for fixed effects. In medicine, such fixed effects would be the age, the health history, and other medical risk factors.

We use the data from the evaluation survey and the grading history of instructors to estimate a statistical model. The model relates the responses in the five categories to question III.5 of the survey to a set of explanatory variables (described below) that identify the fixed effects. The model, known in the literature as an Ordered Multinomial-Choice Model, has been used extensively to study data that take on a finite number of values possessing a natural ordering.¹ Rather than describe the model in technical detail, let us outline some of its key features:

- The model accounts for the fact that students only have 5 choices when responding to question III.5. We can therefore explicitly deal with the distortions caused by this constraint.
- The model can accommodate varying response rates across classes. Importantly, its structure allows us to investigate whether low response rates are related to how the students view the instructor. If absenteeism primarily occurs for other independent reasons (like sickness), the model will quantify the degree to which the precision of our statistical analysis is affected. If absenteeism is not independent, the model will allow us to quantify and correct for the biases caused by non-response.
- The model does not use the average score as the survey measure to be explained. This is important because the average score is a very poor statistical measure of student responses. The average fails to account for the degree of unanimity in the student responses. It also implicitly places uneven weights on individual student responses. (These effects were described in detail in the proposal for this project which is available from the authors.)

The explanatory variables included in the model are the following:

- The student response ratio.
- The course level, taking on a value of 1 if it is a basic course and 0 otherwise.
- A small class size indicator, taking on a value of 1 if it is less than 5 and 0 otherwise.
- A seminar class size indicator, taking on a value of 1 if it is more than 5 and less than 15.
- An early morning indicator, taking on a value of 1 if the class meets before 10 am, 0 otherwise.

- Three variables for the grading reputation of the instructor: The percentage of B grades, the percentage of C grades, and the percentage of D/F grades, given by the instructor in the previous year.

The model estimates are presented in Table 5. The complex structure of the model makes it hard to judge how large an affect a variable has on the scores from simply looking at the coefficient estimates. However, Table 5 does allow us to identify which variables have significant effects and whether they raise or lower scores. In particular, a positive (negative) coefficient estimate implies that an increase in the associated explanatory variable will raise (lower) the numerical score given by a student in each class. The T-statistics in the right-hand column provide a statistical measure of significance. By this measure, all the coefficients except the level of the course are significantly different from zero at the one percent level or less. This means that there is a less than one in a hundred chance of finding the relation we observe in the data between the scores and the explanatory variable if in fact they were unrelated. By this measure, the statistical evidence linking scores to the explanatory variables is extremely strong. Only the level of the course seems unimportant. Teaching a lower or upper level course, *ceteris paribus*, has no systematic effect on the student evaluations.

Table 5: Ordered Multinomial-Choice Model Estimates

Explanatory Variable	Coefficient	Std Dev	T-stat
Response ratio	0.2292	0.0114	20.0951
level 1 course	-0.0216	0.0123	-1.7618
Less than 5	0.0542	0.0133	4.0792
5 to 15	0.0076	0.0118	0.6390
Early morning	0.0309	0.0113	2.7387
B's	-0.0515	0.0117	-4.4060
C's	-0.0840	0.0145	-5.7994
D's	-0.0964	0.0137	-7.0141

The estimated coefficients indicate that higher response rates increase the instructor's score. One interpretation of this finding is that more effective teachers get better attendance at their classes. However, such an interpretation *assumes* that the evaluation forms were distributed in a typical class. If the forms were handed out at a pre-announced review session, students who found the instructor particularly effective may have stayed away because they viewed the review as unnecessary. In this case, a lower score would be associated with a lower response rate because the survey under-represents students who found the instructor particularly effective.

¹ We report estimates for an Order-Logit specification of the model below. We have also estimated specifications within normal (i.e., Order-Probit) and extreme value distributions. In each case, the estimates are very similar to those we report.

Table 5 also shows that teaching a class with fewer than 5 students systematically increases the instructor’s score. This is hardly surprising in view of the personalized instruction students receive in such classes. We also find that teaching early has a positive (rather than negative) and significant effect on student ratings. Could it be that students willing to take early classes are also more receptive to the efforts of their instructors? The last three rows of the table show how the grading reputation of the instructor affects scores. The estimates indicate that a reputation for giving more B’s, C’s, and particularly D/F grades, significantly lowers the score.²

Table VI shows how well this model performs. Based on the model estimates we can make a prediction about the number of students in each class that will answer question III.5 with a check mark in bin 1 though 5. We then compare this prediction against the actual scores. If the explanatory variables explain *all* the differences in scores across classes, our predictions will match the actual numbers. We would then have explained 100 percent of the variation in scores across classes. Alternatively, if the explanatory variables have no effect on scores, we would have explained 0 percent of the variation.

Table 6: Explanatory Power

Bin	Cross-Class Variation
1	4.41%
2	15.98%
3	31.63%
4	67.86%
5	60.57%

Table 6 shows that the model explains over 60 percent of the variation of scores in bins 4 and 5, across all classes. Since most of the teaching evaluation scores are in these bins, this result tells us that the explanatory variables, such as class size, time of day, response rates, and grading reputation, explain more than half of the differences between students marking 4 and 5 in response to question III.5. Alternatively, more than half of the variation of teaching scores of “above average” or “excellent”, has nothing to do with the identity or teaching ability of the instructor!

The model is a highly non-linear one, so how scores change with respect to grading and other characteristics is not obvious. We therefore examine how changing the characteristics of a typical class affect the score. Consider a class in which the response rate is perfect, with size being between 5 and 15, an upper-division class, meeting outside of early morning, and a “grading history” of the instructor of 74 percent A grades, 20 percent B grades, and 6 percent C grades. The first line in Table 7 shows that the predicted score for this class is 4.5.

² We did not include the percentage of A grades since the total percentages add up to one.

Table 7: Changing Class Characteristics

	Response level 1 Rate	0.0	less than 5	5 to 15	Early Morning	B%	C%	D/F%	Predicted Score
Initial Characteristics	1.0	0.0	0.0	1.0	0.0	20	6	0	4.50
Less response	0.8	Same	Same	Same	Same	Same	Same	Same	4.35
Lower level	Same	1.0	Same	Same	Same	Same	Same	Same	4.49
Less than 5	Same	Same	1.0	Same	Same	Same	Same	Same	4.72
More than 15	Same	Same	Same	0.0	Same	Same	Same	Same	4.50
Early morning	Same	Same	Same	Same	1.0	Same	Same	Same	4.54
More B's	Same	Same	Same	Same	Same	30	Same	Same	4.41
More C's	Same	Same	Same	Same	Same	Same	16	Same	4.47
More D's	Same	Same	Same	Same	Same	Same	Same	10	4.37
Lower Response Rate and University Grade Guide	0.8	1.0	Same	Same	0.0	54	13	1	4.22
Lower Response Rate and Lower than University Grade Guide	0.8	1.0	Same	Same	0.0	50	13	6	4.16

The middle panel of Table 7 shows how the predicted score changes when we vary each of the explanatory variables in turn. The effects of changing each variable in isolation are generally quite small. The lower panel of the table shows what happens if the instructor has a response rate of 80 percent, and followed the university grading guidelines in the previous year; giving 32 percent A grades, 54 percent B grades, 13 percent C grades, and 1 percent D/F grades. Here we see that the predicted score falls from 4.5 to 4.22. While this may not seem particularly striking in numerical terms, it has a dramatic effect on the relative ranking of the instructor. A score of 4.5 places the instructor in the upper 52nd percentile of scores across the university, i.e., just above the median. A score of 4.22, places the instructor among the lowest 31 percent. Thus, an instructor's decision to abide by the university grading guideline would push the instructor's score from around the university average to one in the lower third. If the instructor adopted a somewhat stiffer grading policy,

shown in the last row, the score would fall to 4.16. At this level, the instructor's score falls in the lowest 27 percentile of the university distribution.

The message from these experiments is clear. If a typical instructor decides to adopt the stiffer grading standard described by the university's guideline, the instructor can expect to see a significant fall in evaluation scores independently of any teaching ability displayed in the classroom. While the actual fall in scores is not particularly large, it is large enough to move an instructor from an average to a well-below average score judged by university norms. Any instructor not wishing to risk being regarded as a "poor" teacher would clearly think twice before following the universities grading guidelines. Perhaps the rarity of low grades at Georgetown says as much about the faculty and the incentives they face, as it does about the academic quality of the students.

Identifying Effective Teaching

To this point we have focused on the effects of class characteristics and grading reputation. We now examine the central issue of whether the evaluation survey provides any reliable guide to the teaching effectiveness of individual faculty.

To address this issue, we re-estimate our model with additional variables that identify the instructor for each course. If the instructor is no more or less an effective teacher than any other faculty member, the variable identifying the instructor will not "explain" any of the student scores in classes he or she taught. In this case, the cross-class difference between the scores for this class and any other are solely attributable to characteristics of the class and the instructor's grading reputation – they have nothing to do with what the instructor actually did in the classroom. Alternatively, if the instructor was highly effective, we should find that the instructor variable has a positive and significant effect on the student scores. In this case, part of the cross-class difference in scores is attributable to the instructor. By similar reasoning, the instructor variable will have a significant negative effect on the student scores in cases where instruction has not been effective.

To make this procedure operational, we must decide on what constitutes a "significant" effect. We follow standard statistical practice by reporting our results for several different levels of confidence. For example, using the 99 percent confidence level, we only classify an individual as being an ineffective teacher if there is a less than 1 percent chance that the student scores were entirely independent of the instructor's performance in the class. Similarly, using the 95 percent confidence level, we only classify an individual as being highly effective if there is less than a 5 percent chance that the student scores were entirely independent of the instructor's classroom performance. Thus, one minus the confidence level measures the probability of making an erroneous assessment of an individual's teaching effectiveness.

Clearly, opinions may differ as to the appropriate confidence level to use in making a particular decision related to the teaching effectiveness of individual faculty. Opinions may

also differ on the confidence level appropriate for different decisions related to an individual's teaching effectiveness. For example, many may argue that the confidence level appropriate for a tenure decision should be higher than for an annual merit review. Similarly, the confidence level appropriate for judging low teaching effectiveness may differ from the level for judging high effectiveness. To make our findings applicable to as wide an audience as possible, Table 8 reports results for confidence levels ranging from 75 to 99 percent.

Table 8: Identifying Teaching Effectiveness

A: Low Effectiveness						
Mean Score Range	Number	Confidence Levels				
		99%	97.5%	95%	90%	75%
0.0 - 5.0	646	19.8	17.7	25.9	29.7	37.5
0.0 - 4.0	115	88.7	92.2	94.5	95.1	98.3
4.0 - 5.0	531	4.9	7.7	10.9	15.4	24.3
4.8 - 5.0	119	0.0	0.0	0.0	0.0	0.0
4.6 - 4.8	131	0.0	0.0	0.0	0.0	0.8
4.4 - 4.6	112	0.9	0.9	0.9	0.9	6.3
4.2 - 4.4	98	5.1	10.2	21.4	31.6	64.3
4.0 - 4.2	71	28.2	42.3	50.7	69.0	81.7
3.8 - 4.0	33	67.7	81.8	90.9	90.9	97.0
3.6 - 3.8	28	96.4	96.4	96.4	96.4	100.0
3.4 - 3.6	12	100.0	100.0	100.0	100.0	100.0
B: High Effectiveness						
Mean Score Range	Number	Confidence Levels				
		99%	97.5%	95%	95%	97.5%
0.0- 5.0	646	20.1	17.7	21.0	30.2	39.8
0.0 -4-0	115	00	0.0	0.0	0.0	0.0
4.0 - 5.0	531	24.5	27.5	32.2	36.7	48.8
4.8 - 5.0	119	56.3	63.0	72.3	75.6	80.7
4.6 - 4.8	131	35.9	41.2	49.6	58.0	84.0
4.4 - 4.6	112	11.6	12.5	15.2	21.4	35.7
4.2 - 4.4	98	2.0	2.0	2.0	4.1	9.2
4.0 - 4.2	71	1.4	1.4	1.4	1.4	2.8
3.8 - 4.0	33	0.0	0.0	0.0	0.0	0.0
3.6 - 3.8	28	0.0	0.0	0.0	0.0	0.0
3.4 - 3.6	12	0.0	0.0	0.0	0.0	0.0

Table 8 shows how the mean scores from Question III.5 relate to the teaching effectiveness of individual faculty for a range of confidence levels. The left hand columns of the table break the mean scores received by each faculty member into different ranges and show how many faculty fall into each range. Panel A reports the fraction of the faculty within

a particular mean score range with a significantly low teaching effectiveness. Panel B reports the fraction of the faculty with significantly high teaching effectiveness.

The results in Panel A are striking. They quantify exactly how difficult it is to identify ineffective teaching from the mean survey scores with any reasonable level of accuracy. To illustrate this point, consider the first row in the panel. Of the 646 faculty in our study, 19.8 percent displayed low teaching effectiveness at the 99 percent confidence level. While this is close to the 19 percent of the faculty with mean scores of less than 4.0, it does not mean that 4.0 can act as a reliable cutoff point. Since only 88.7 percent of those scoring less than 4.0 display significantly low effectiveness, more than 11 percent of this group would be incorrectly identified as ineffective teachers. This implies an error rate of $1 - 0.887 \times 0.99$, which is approximately 12 percent. Matters are even worse if we use the 75 percent confidence level. Here the error rate is over 26 percent. In fact, the error rate associated with the 4.0 cutoff remains above 10 percent for any level of confidence level.

The 4.0 cutoff is also hard to justify on equity grounds. For example, at the 75 percent confidence level, 24.5 percent of faculty scoring above 4.0 display significantly low effectiveness. These individuals appear above the cutoff because their grading reputations and class characteristics counteract their ineffectiveness in the classroom. It is hard to see why this group should be viewed as more effective teachers than individuals scoring less than 4.0. In sum, these results show that a cutoff score of 4.0 for ineffective teaching cannot be justified on either accuracy or equity grounds irrespective of the confidence level one may choose to use.

The lower portion of panel A shows the relation between teaching effectiveness and survey scores over narrower ranges. These results provide some guidance how to choose a cutoff point. Consider, for example, a cutoff score of 3.8. From the table we see that 96.4 percent of instructors scoring between 3.6 and 3.8 display low effectiveness at the 99 percent confidence level. The error rate for this group is therefore $1 - 0.964 \times 0.99 = 0.046$. This is lower than the error rate associated with any other confidence level but it is well above the 1 percent error rate that is typically tolerated in important decision-making. To achieve this lower error rate, the cutoff score needs to be 3.6.

Identifying high teaching effectiveness from the survey is just as difficult. Panel B of Table 8 reports the fraction of faculty with significant effectiveness for different scoring ranges and varying confidence levels. Approximately 20 percent appear highly effective teachers at the 99 percent confidence level. This closely matches the fraction that are ineffective. Unsurprisingly, none of the faculty scoring below 4.0 are highly effective at any confidence level. What is more surprising is that so few of the faculty scoring over 4.0 are effective. If we classified everyone with scores above 4.0 as highly effective, the error rates range from 63 to 76 percent. Moreover, the error rates do not decline significantly if we raise the cutoff point. The lowest error rate for effective teaching is still 31 percent and comes from a cutoff score of 4.8.

In our view, the results in Table 8 completely undermine the current methods for identifying effective and ineffective teaching. The use of a 4.0 cutoff score to signal ineffective teaching results in so many erroneous judgments that it must be viewed as essentially arbitrary. Similarly, one cannot identify particularly effective teaching from high survey scores without

making a lot of errors. While there are undoubtedly very effective teachers at Georgetown, there is no evidence that these individuals can be reliably identified from the student surveys.

Conclusions

This empirical analysis of student survey data shows how difficult it is to accurately evaluate “teaching effectiveness”. But it would be a mistake to draw from this study that the current survey instrument needs to be replaced by another instrument, with more differentiated and probing questions, but with the same numerical scoring device. We are confident that we would be able to replicate similar error rates for assessing teaching from another instrument, which follows a similar *modus operandi* as the current instrument.

One should conclude from this study that teaching evaluation data should only be used cautiously, to separate “ineffective” teachers from “effective teachers” with a cut-off at a very low range, of 3.6. Only at this rate can one be reasonably sure that one is not unfairly categorizing good teachers as “ineffective”. Arbitrary scores of 4.0 or above, as cut-offs for promotion or tenure, are just that.