During the past decade or so, many colleges have placed an increasing emphasis on “teaching effectiveness” in faculty promotion, tenure, and salary reviews. In most cases, the mechanism used to measure teaching effectiveness is a locally developed evaluation form that is completed by an instructor’s students toward the end of the course, usually before students have received their final course grades. The practice of using student evaluations of teaching (SETs) to evaluate faculty teaching effectiveness raises a number of concerns, including the basic validity of these forms and their sensitivity to external biases. The question of validity involves the extent to which SETs (or items on these forms) accurately predict student learning. Questions of bias involve the possibility that student responses are influenced by factors unrelated to the faculty member’s instructional effectiveness. The topic of this article is the biasing effect that faculty grading practices have on SETs. A broader discussion of this and related issues may be found in my book The GPA Myth, from which most of the following analyses are drawn.

Both the validity of SETs and potential biases to SETs have been discussed extensively in the educational literature. A simple search of the ERIC database produces thousands of articles concerning various aspects of SETs, and Greenwald summarizes more than 170 studies that examined the specific issue of whether SETs represented valid measures of student learning. Clearly, a comprehensive review of this literature is not possible here, and so I will simply summarize the current state of knowledge concerning the relation between grading practices and student evaluations of teaching by asserting that it is rather confused. Much of this confusion arises from the fact that it is difficult to conduct experiments to investigate this relationship because of constraints involving human subjects. Those experiments that have been conducted tend to confirm the notion that higher student grades cause higher teacher course evaluations, but over the past 2 or 3 decades several other explanations for the correlation between grades and...
Survey Instrument

Listed below is a subset of the items that appeared on the DUET Web site.

The instructor’s concern for the progress of individual students was
1) Very Poor 2) Poor 3) Fair 4) Good 5) Very Good 6) Excellent
7) Not Applicable

How effective was the instructor in encouraging students to ask questions and express their viewpoints?
1) Very poor 2) Poor 3) Fair 4) Good 5) Very Good 6) Excellent
7) Not Applicable

How would you rate this instructor’s enthusiasm in teaching this course?
1) Very bad 2) Bad 3) Fair 4) Good 5) Very Good 6) Excellent
7) Not Applicable

How easy was it to meet with the instructor outside of class?

How does this instructor(s) compare to all instructors that you have had at Duke?
1) Very bad 2) Bad 3) Fair 4) Good 5) Very Good 6) Excellent 7) Not Applicable

How good was the instructor(s) at communicating course material?
1) Very bad 2) Bad 3) Fair 4) Good 5) Very Good 6) Excellent 7) Not Applicable

To what extent did this instructor demand critical or original thinking?
1) Never 2) Seldom 3) Sometimes 4) Often 5) Always 6) Not Applicable

How valuable was feedback on examinations and graded materials?
1) Very poor 2) Poor 3) Fair 4) Good 5) Very Good 6) Excellent
7) Not Applicable

How good was the instructor at relating course material to current research in the field?
1) Very bad 2) Bad 3) Fair 4) Good 5) Very Good 6) Excellent 7) Not Applicable

Experimental Design

The data used in this investigation were collected during an experiment conducted at Duke University during the fall semester of 1998 and the spring semester of 1999. The experiment, called DUET (Duke Undergraduates Evaluate Teaching), utilized a commercially operated Web site to collect student opinions on the courses they either had taken or were taking. The DUET Web site was activated for two 3-week periods, the first beginning the week prior to fall registration in 1998 and the second a week prior to spring registration in 1999. Both periods fell approximately 10 weeks into their respective semesters. The survey was conducted immediately before and during registration as an incentive for students to participate; by completing the DUET survey prior to registration, students could view course evaluation data and grade data for courses they planned to take the following semester. An important aspect of the experimental design involved the manner in which data were collected for first-year students. Participating first-year students completed the survey for their fall courses twice: once before completing their fall courses and receiving their final grades, and once after. Because one DUET survey item asked students what grade they received or expected to receive in their courses, the responses collected from freshmen at the two time points provide an ideal mechanism for investigating the influence that expected and received student grades have on student evaluations of teaching.

Upon entering the DUET Web site, students were initially confronted with text informing them that course evalu-
ation data collected on the site would be used as part of a study investigating the feasibility of collecting course evaluations on the Web. They were also told that their responses would not be accessible to either their instructors, faculty not associated with the experiment, other students, or university administrators. After indicating their consent to participate in the study, items on the survey were presented to students in groups of five to seven questions. With only one or two exceptions, all items included a “Not Applicable” response. Thirty-eight items were included on the survey, but, because of space constraints, attention here is restricted to nine questions that relate to student-instructor interaction. These items and possible student responses appear in the sidebar on the previous page.

During the two semesters that the DUET Web site was active, 11,521 complete course evaluations were recorded. Each evaluation consisted of 38 item responses for a single course. Of the 6,471 eligible full-time, degree-seeking students who matriculated at Duke University after 1995, 1,894 students (29 percent) participated in the experiment. Among first-year students, approximately one-half participated in both the fall and the spring surveys. A detailed investigation of the relationship between various demographic variables and response rates of the corresponding groups is provided in The GPA Myth. That analysis suggests that response patterns were not significantly affected by the variables most associated with differential response rates. In particular, although response rates did vary according to students’ gender, ethnicity, academic year, academic major, and GPA, student response patterns did not vary substantially according to these categories.

Statistical Analysis

Because student responses to the DUET survey take the form of ordinal data—that is, responses fall into ordered categories—it is natural to estimate the effects of received and expected grades on SETs using ordinal regression models. To this end, I assume the existence of an unobserved, or latent, variable that underlies each student’s perception of an item. As the value of this latent variable increases, so too does the probability that the person rates the item in one of the higher categories. This concept is illustrated in Figure 1. The curve at the top of the figure represents the distribution of the latent variable for an individual who is likely to strongly disagree or to disagree with a statement. Individuals whose latent variables are drawn from the distribution in the middle of the figure are more likely to respond in a neutral way, while individuals whose latent distributions are depicted at the bottom of the figure are likely to agree or strongly agree with the question.

Statistical issues that arise in analyzing ordinal data using latent variable models involve establishing a scale of measurement (i.e., establishing the location of 0 on the horizontal axis in Figure 1), estimating the locations of the category cutoffs (the vertical, dashed lines in Figure 1 that separate the response categories), and, most important, modeling
the effects that explanatory variables have in shifting the distribution of the latent variables. The first problem—

establishing a scale of measurement—

can be solved by setting the shape of the curves used to describe the distributions of the latent variables and by pinning down the value of one of the category cutoffs. For this purpose, in the analyses that follow, the shape of these distributions is assumed to be a standard logistic density function, and the value of the lowest category cutoff is fixed at 0.

To model the location of the logistic curve that determines the probability that a particular student responds in a certain way on a given DUET item (i.e., to model the location of the center of the curves depicted in Figure 1), assume that for each DUET item and student, there is a value, say $c_i$, that represents the student's perception of the given DUET item for the course in question, excluding influences of expected or received grade. That is, $c_i$ is defined to represent the central value of student's latent perception distribution for that DUET item when no consideration is given to expected or received grade. For brevity, call this value the course effect.

For DUET data collected in the fall of 1998, the effects of expected grade on students' perception of items are modeled by assuming that the influence of expecting a grade of, say, $g$, on the $i$th student's perception of a given DUET item is to shift the latent perception distribution of that student by an amount equal to $E_g$. Student expectations of their grades were defined using their actual course grades for their fall courses. For this reason, modifications to students' perceptions of their fall courses during the spring survey period are attributed to received grade rather than expected grade. Let $R_g$ denote the influence that receiving a grade of $g$ has on a student's perception of a course attribute in the spring, where $g$ is coded as before but now represents the grade actually received in the course. With this notation in hand, the central value of the curve describing the $i$th student's perception of a DUET item, including possible influences of expected grade, was thus assumed to be $c_i + E_g + R_g$.

Assumptions made for data collected in the spring of 1999 are similar. As in the fall, $c_i$ denotes the central value of the $i$th student's latent perception distribution for a given DUET item, where, as before, influences of expected or received grades are excluded. The relative values of $c_i$ are assumed to be constant across semesters for the same student/DUET item combinations.

An important difference between data collected in the fall and spring is that students who completed the survey in the spring had, by that time, received their actual course grades for their fall courses. For this reason, modifications to students' perceptions of their fall courses during the spring survey period are attributed to received grade rather than expected grade. Let $R_g$ denote the influence that receiving a grade of $g$ has on a student's perception of a course attribute in the spring, where $g$ is coded as before but now represents the grade actually received in the course. With this notation in hand, the central value of the curve describing the $i$th student's perception of a DUET item, including possible influences of expected grade, was thus assumed to be $c_i + E_g + R_g$.

Finally, distinct category cutoffs were defined separately for the fall and spring data. By so doing, systematic changes in student perceptions of classes over time were naturally integrated into the probability model for the student responses.

---

### Table 1—Estimated Effects of Expected Grades on DUET Items.

<table>
<thead>
<tr>
<th>Item</th>
<th>$E_D$</th>
<th>$E_C$</th>
<th>$E_{C_0}$</th>
<th>$E_{C_1}$</th>
<th>$E_{C_2}$</th>
<th>$E_{C_3}$</th>
<th>$E_{C_4}$</th>
<th>$E_{C_5}$</th>
<th>$E_{C_6}$</th>
<th>$E_{C_7}$</th>
<th>$E_{C_8}$</th>
<th>$E_{C_9}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructor concern</td>
<td>1.53</td>
<td>0.07</td>
<td>0.93</td>
<td>2.62</td>
<td>2.85</td>
<td>3.02</td>
<td>2.70</td>
<td>3.13</td>
<td>3.57</td>
<td>4.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encouraged questions</td>
<td>2.98</td>
<td>1.07</td>
<td>1.61</td>
<td>2.37</td>
<td>2.44</td>
<td>2.90</td>
<td>3.09</td>
<td>3.26</td>
<td>3.53</td>
<td>3.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructor enthusiasm</td>
<td>2.99</td>
<td>0.67</td>
<td>2.12</td>
<td>2.70</td>
<td>3.19</td>
<td>3.50</td>
<td>3.28</td>
<td>3.44</td>
<td>3.96</td>
<td>4.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructor availability</td>
<td>2.24</td>
<td>2.67</td>
<td>2.10</td>
<td>2.85</td>
<td>3.27</td>
<td>3.35</td>
<td>3.06</td>
<td>3.35</td>
<td>3.60</td>
<td>3.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructor rating</td>
<td>2.11</td>
<td>-0.43</td>
<td>1.99</td>
<td>2.34</td>
<td>2.92</td>
<td>2.71</td>
<td>2.77</td>
<td>2.96</td>
<td>3.36</td>
<td>4.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructor communication</td>
<td>2.78</td>
<td>0.52</td>
<td>1.97</td>
<td>1.54</td>
<td>2.88</td>
<td>3.30</td>
<td>3.29</td>
<td>3.52</td>
<td>4.13</td>
<td>4.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical thinking</td>
<td>1.26</td>
<td>2.10</td>
<td>2.69</td>
<td>1.51</td>
<td>2.63</td>
<td>3.14</td>
<td>3.39</td>
<td>3.13</td>
<td>3.04</td>
<td>3.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usefulness of exams</td>
<td>1.70</td>
<td>0.33</td>
<td>0.88</td>
<td>2.40</td>
<td>2.69</td>
<td>3.04</td>
<td>3.20</td>
<td>3.49</td>
<td>3.53</td>
<td>4.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Related course to research</td>
<td>0.60</td>
<td>-0.29</td>
<td>1.10</td>
<td>2.14</td>
<td>2.30</td>
<td>3.01</td>
<td>3.11</td>
<td>3.42</td>
<td>3.63</td>
<td>4.15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---
To summarize this model for the paired student responses, we have assumed that student responses to each DUET survey item depend first on an underlying perceptual variable that determines each student’s general impression of the course attribute in question and second on the student’s expected or received grade in that course. A possible drawback of this formulation is that, at first glance, there appear to be almost as many parameters in the model as there are paired student responses to the survey. Indeed, if only the fall or the spring data were examined separately, there would be more parameters in the model than observations in the data, since each student response to each DUET item for each course invokes its own value of \( c_i \). However, with two semesters of data, there were two observations of each student’s response to each DUET item for each course taken, and only one value of \( c_i \) to estimate for each such pair. For purposes of estimating the remaining model parameters, this results in a surplus of observations in number equal to the response pairs. Because more than 2,500 paired responses were obtained for each item, this leaves ample data available for estimating the fixed effects of grades.

An advantage of estimating \( c_i \) for each paired student response is that it permits the effects of grades on student evaluations of teaching to be separated from course attributes and other student background variables. Presumably, prior student interest, student motivation, academic prowess, and the intellectual level at which an instructor pitched his or her lectures are all absorbed in this variable. The consequence of fitting a random course effect for each paired response is that estimates of grade effects are based only on changes to survey responses that accompany expected and received grades.

Posterior means of the expected grade and received grade parameters estimated from this model, obtained using estimation procedures described in *Ordinal Data Modeling*, are provided in Tables 1 and 2. Posterior standard deviations are indicated in parentheses below each estimate. The coefficients in Tables 1 and 2 can most easily be interpreted in terms of odds and odds ratios. Because the latent student perception variables were assumed to have standard logistic distributions, the ratio of the probability that a student rated an item above a given category, say \( m \), to the probability that he or she rated it less than or equal to category \( m \) in the fall 1998 survey can be expressed

\[
\Pr(\text{response} > m) \exp(-\gamma_m + c_i + E_g). 
\]

In this expression, \( \gamma_m \) denotes the category cutoff for category \( m \) (illustrated in stylized form by a vertical dashed line in Figure 1), and \( E_g \) denotes the coefficient listed for an expected grade of \( g \) in Table 1. This ratio of probabilities is called the odds.

The corresponding odds for the probability that a student responded in category \( m \) or higher during the spring 1999 survey is given by

\[
\frac{\Pr(\text{response} > m)}{\Pr(\text{response} \leq m)} \exp(-\gamma_m + c_i + R_g). 
\]
where $R_i$ is the effect of a received grade of $g$, as listed in Table 2.

Such odds depend on specific course effect parameters $c_i$ and the category cutoff parameter $\gamma_m$. However, the impact of grades on student responses to the DUET items can be examined more directly by calculating the odds ratios for different received or expected grades. The odds ratios do not depend on either the course effects $c_i$ or category thresholds $\gamma_m$. For example, suppose interest focuses on the differential effect that a student expectation of a grade of A versus a grade of B has on a student's response to the question regarding instructor concern.

Thus, the ratio of the odds that a student expecting an A rates this item higher than (any) category $m$ to the same odds for a student expecting a B is 1.75.

For the spring data, received rather than expected grades are relevant. Again considering the item involving instructor concern, the corresponding odds ratio is

$$
\exp(R_A - R_B) = \exp(3.45 - 2.23) = 3.10.
$$

In other words, the odds that a student gives an instructor a “high rating” on concern is increased by a factor of more than 3 if the student receives an A instead of a B!

Figure 2 provides an illustration of the coefficients listed in Table 1. In the plots depicted in this figure, the posterior means of the expected grade coefficients for the nine DUET items listed in the sidebar are displayed as a function of increasing expected grade. One striking aspect of the plots in Figure 2 is the nonlinearity of the trend at the lowest grade levels. Apparently, students who expect to receive a grade
of D+ or lower are less likely to attribute their poor performance to the instructor than are students who receive grades in the C range. A plausible explanation for this phenomenon, which manifests itself for most of the items in the other groups as well, is that students who expect to receive grades of D+ or lower have not adequately invested themselves in their courses. Keeping in mind that fewer than 3 percent of all grades assigned to first-year students at Duke are lower than a C-, and that essentially any level of student commitment would earn a grade of at least a C- in a vast majority of Duke courses, it is likely that students who expect to receive such low grades recognize that their poor performance is due to a lack of effort. In other words, students may be willing to attribute poor performance to themselves if such performance results from a lack of effort rather than lack of ability.

Even more striking than the nonlinearity of the trend in the grade effects at the lowest grade levels is the monotonicity of this trend at more normal grade levels. For eight out of the nine questions in this group, the tendency of students to rate more highly those courses for which they received higher grades is nearly uniform. The exception occurs for the item involving the extent to which the instructor required critical thinking, which levels out above grades of B. Because this is the only question in this group that tangentially involves course difficulty, the explanation for this deviation might involve a balancing of the presumably negative effect of the use of stringent grading as a tool to require critical thinking and the positive, attributional effect of receiving higher marks.

Plots comparable to those in Figure 2, but based on received grade, are displayed in Figure 3. Qualitatively, the two sets of plots are commensurate. One feature of these plots is, however, more pronounced in the plots based on received grade: the differences that accompany changes in student evaluations between students receiving a C+ and a B-. For many items, this difference results in a notable spike in the otherwise dominant trend. This spike is likely caused by the relatively large perceptual difference among undergraduates between the value of a C grade and a B grade.

The tendency of students to rate items more highly upon receipt of a higher grade cannot easily be explained by the prevalent theories that have been posed to explain observed correlations between grades and student evaluations of teaching. The experimental design used to measure these trends essentially eliminates the possibility that these effects might have been caused by teacher effectiveness or most intervening variables. The simplest and most reasonable explanation for the trends reported in Figures 2 and 3 is that grade expectations or received grades caused a change in the way students perceived their teacher and the quality of instruction. The extent to which these data support alternative explanations for the
As an increasing number of universities use student evaluations of teaching in administrative decisions that affect the careers of their faculty, the incentives for faculty to manipulate their grading policies in order to enhance their evaluations increase.

correlations of grades and SETs is discussed in The GPA Myth.

Summary

Data collected during the DUET experiment provides a unique glimpse into the causal nature of student grades on teacher-course evaluations. Although any number of extraneous factors may act to influence both grades and SETs, the analysis of the DUET data suggests that there is a direct effect of grading policy on SETs beyond what can be explained by such extraneous factors.

A unique feature of this analysis was that two student responses were obtained for each student-course combination. The first response was obtained while students were still taking their courses and before they had received their final course grades. The second response was collected after students had completed their courses and after they had received their final course grades. By contrasting the two responses obtained from each student, highly significant, substantively important effects of student grade were discovered on all of the DUET items. These findings corroborate the findings of earlier grade-manipulation studies and a preponderance of correlational studies. Because the design of the DUET experiment effectively eliminated the possibility that unobserved environmental factors were responsible for these effects, the results from this analysis provide conclusive evidence of a biasing effect of student grades on student evaluations of teaching.

From a policy viewpoint, the findings of this study are important. As an increasing number of universities use student evaluations of teaching in administrative decisions that affect the careers of their faculty, the incentives for faculty to manipulate their grading policies in order to enhance their evaluations increase.

References and Further Reading


