Estimating the Impact of Relative Expected Grade on Student Evaluations of Teachers

Andrew Ewing *
Department of Economics,
University of Washington

February 13, 2009

Abstract

Grade inflation over the past few decades has been a concern for many universities. Course evaluation scores are known to be positively correlated with students' expected grades, and this paper tests whether or not there is an incentive for the instructor to "buy" higher evaluation scores by inflating grades. To test this hypothesis, I use unique data from the University of Washington's Office of Educational Assessment that includes a measure of each student's relative expected grade in the course. I find that there is an incentive to inflate grades even after accounting for the potential endogeneity of the relative expected grade variable due to unobserved teacher productivity and after accounting for the unobserved heterogeneity of instructors and departments by using a fixed effects estimation. In my estimations, department and instructor fixed effects account for a significant part of the measured effect of relative expected grade on evaluations, and by not including them, the resulting estimated impact of relative expected grade on evaluations is biased upwards. This suggests that adjustments to evaluations for possible grade inflation need to be done on a departmental basis, and not by a university-wide average.

*email: ewingam@u.washington.edu. PRELIMINARY–Comments welcome. I would like to thank the members of my committee: Shelly Lundberg, Gene Silberberg, Levis Kochin, Mark Long, and Marge Plecki for their support. I would also like to thank Robin Anderson, Byron Tsang, Katie Ewing, Joseph Calhoun, the participants at the University of Washington Labor/Development Brownbag Seminar Series, and the participants at the 2008 Southern Economic Association Conference for helpful comments. Finally, I would like to thank Deb McGhee from the UW OEA and Susana Angkico from the UW Registrar's Office for providing data.

1 Introduction

At many universities, there is a strong positive correlation between students' expected grades in a college course and their evaluations of the instructor for that course. One suggested reason for this correlation is that if a student expects to fail the course, she most likely will not give an "excellent" evaluation of the teacher, regardless of whether or not that teacher really is excellent; if a student expects to get an "A" in the class, she may be nicer on the evaluation form. This relationship presents a problem since one of the major factors in deciding on promotions and merit pay raises in higher education is the quality of student teaching evaluations. (Hostetler et al., 2004) This implies that instructors may have an incentive to inflate grades to get higher evaluations from their students.

This paper focuses on using several econometric techniques to analyze the link between expected grades and students' evaluations of teachers (SETs) and to correct for effects that may bias results. Namely, I use both a two-stage least squares/instrumental variables approach and a fixed effects approach as two methods of controlling for unobserved instructor characteristics. There are pros and cons to using the different approaches that will be discussed; however, an important advantage in the fixed effects approach is the ability to capture structural differences between departments. In particular, department effects on grading policies have yet to be addressed in the literature. The end result of this study is a range of estimates that attempt to measure the impact of grading leniency on SETs that control for both department and instructor characteristics. These measurements are beneficial to the college or university as they consider the issue of grade inflation and decide whether or not to adjust instructors' final SET scores.

It should be noted that there are several measures that are used for grades in previous studies' estimations. Some authors use actual grades in their models, whereas some use expected grades. While each measure has its own merits, the seemingly preferred measure in the literature is expected grade. Students do not know their actual grade at the time of the evaluations, but we assume they have some expectation of the grade they will receive.

Questions have been raised about how accurately actual grades measure expected grades, as students tend to be irrationally overconfident in their grade expectations. (Nowell and Alston, 2007) Therefore, expected, and not actual, grade is a determinant of SETs.

Accepting expected grade as the preferred measurement, the measurement can be refined further by using a relative measure of expected grades instead of an absolute level. That is, the preferred measure, in terms of representing a determinant of SETs, is the difference between a student's expected grade and what that student is accustomed to receiving. (Isely and Singh, 2005) For example, a student that enters the course with a 3.8 cumulative GPA may be upset if she receives a grade of 3.5, while a student who has a 2.8 cumulative GPA may be pleased with the same absolute grade of 3.5. Unfortunately, researchers typically calculate this measure of "relative expected grade" after the course is over by using the actual cumulative GPAs of the students at the section level. This raises selection concerns since evaluations are typically anonymous and some students simply do not fill out the forms. If there is a systematic difference between those students that do fill out evaluations and those that do not, this version of relative expected grade that is calculated after the course is finished may contain measurement error due to selection bias.

To the best of my knowledge, my dataset is unique in that it includes a question on the evaluation form that directly asks students how their expected grade relates to what they *individually* are accustomed to receiving. By asking this question directly to the student, my dataset bypasses the above selection bias problem. This gives a direct indication of how an individual student's relative expected grade affects her evaluation of her instructor.

I find that after using a two-stage least squares/instrumental variables approach to account for the potential endogeneity of the relative expected grade variable due to unobserved teacher productivity, the estimated impact of relative expected grade on evaluations is large and significant. In other words, my results suggest that there remains an incentive for instructors to grade leniently even after instrumenting for the effect of teacher productivity. When estimating the model with instructor and department fixed effects, I find that the incentive to grade leniently remains significant but of a smaller magnitude than the two-stage least squares/instrumental variables estimates. If these fixed effects are not included, the estimated impact of relative expected grade on SETs is biased upwards. This suggests that if a university were to adjust evaluations for possible grade inflation, it needs to account for these unobserved department and instructor characteristics.

2 SETs and the "Principal-Agent-Client" Problem

Students' evaluations of teachers as measurements of instructional quality remain popular at universities and colleges, but they are riddled with potential biases. This leads to the question, "Why not just abandon the student evaluation system altogether?" This most likely will not happen since SETs are viewed as the lowest cost option for colleges and universities to get measurements of instructional quality. There is no payment given to the student that fills out the form and thus the only costs to the college involve the collection, processing, and analysis of the data.[1]

The problems associated with SETs in higher education fall under the umbrella of the broader "principal-agent-client" problem. In general, introducing hierarchies into an agency problem increases information and monitoring costs. (Garicano, 2008) In this example, the principal (department/college) introduces an evaluation instrument (SET) that is to be used by the clients (students) in order to gather information about the agent's (instructor's) performance. A problem arises given that the instructor has two methods of achieving better SET scores. The department would prefer that the instructors improve SET scores by actually improving their teaching, but instructors may find it less costly to simply inflate grades. The issue of a principal finding an appropriate evaluation instrument for the agent

[1] One alternative method for evaluating teachers would be peer evaluations (which can also be used as a complement to SETs). There are several drawbacks to this option. Often, these peer observations are announced in advance, possibly masking the effectiveness of an instructor over the span of a semester or quarter. In fact, the instructor may even be able to choose her particular session of a class for evaluation and will strategically choose one that showcases her best qualities. SETs have the advantage that the students have the opportunity to attend every class. If each and every class were to be evaluated by a peer, the costs would be extremely prohibitive.

that does not also introduce perverse incentives lies at the heart of Personnel Economics.[2]

2.1 Students–The Clients

It is important to note that the students' completion of the SET questionnaire is anonymous and voluntary. In other words, when a student completes the evaluation form, she is essentially helping provide a public good at a small, positive cost to herself. Why would a student engage in this behavior? Whatever the reason, the average response rate[3] in this particular dataset is 68%.

Another question that arises is of accuracy. How accurately and fairly do students respond to the questions given that they typically take only a few minutes to fill out the form. Marlin (1987) asked this very question on a survey of students at Western Illinois University and Appalachian State University and found that over 90% of those surveyed claimed that they are almost always or at least most of the time "fair and accurate...and give adequate thought and effort to the rating process." Noting that students behave in this fashion does not indicate that grading leniency does not have an impact on SETs, just that students perceive themselves as acting fairly.

Presumably, students take courses for the learning that is produced. Part of this learning is due to student effort in the class and the other part is due to the instructor's contribution. Whatever the productive combination may be, overall learning in the course is ubiquitously measured by the students' grades in the course. Therefore, students will value the grade they get in the class in and of itself since it is an outward indication to others of what they have learned in that class. It could be the case that students who achieve high grades (potentially through lenient grading) believe that they learned more, and reward their instructors by giving higher SET scores. But again, the value does not lie just in the absolute value of the grade but in the grade relative to what they as an individual student are accustomed to

[2]See Chapter 8 of Lazear (1995) for an overview of issues in evaluation.

[3]Calculated as the ratio of the number of questionnaires turned in to the number of students enrolled in the class.

receiving.

As far as the answers given on the SET form are concerned, students are able to evaluate the instructor and the course given all of the information above except for their actual grades. Since the SET form is typically handed out with a portion of students' grades yet to be determined, students answer the questions armed with only their *expected* grade.

2.2 Instructors–The Agents

Different instructor characteristics will be important determinants of their behavior as far as their attitudes to improving their evaluation scores. In particular, the rank of the instructor will be important since the different ranks of instructors will be important in department deliberations over promotion, tenure, and merit pay raises. For example, Teaching Assistants (TAs) and Lecturers may have more incentive to inflate grades since evaluations are more closely linked to continuance of funding. In addition, since there is less or no emphasis on research for Lecturers, evaluations are the primary measure of quality.

The rank of the instructor is also a good indicator of her experience teaching a variety of courses and, in particular, the same course more than once. Instructors who have taught the same course several times face significantly fewer costs (e.g., lecture preparation, test banks, etc.) than new instructors. Each time an instructor teaches a course, though, she expends some positive amount of effort and therefore incurs costs. The costs come in several different forms, from the actual time spent teaching and grading to the opportunity cost associated with lost research productivity. This cost of effort depends on her own ability as a teacher and the quality of the students in her class.

With that said, the instructor may also value the learning produced, whether it enters directly into the instructor's utility function or if the learning creates other benefits in the long run. For example, if the instructor is instrumental in preparing a particular student for graduate school, the student may reciprocate by recommending that instructor after the class is finished.

The instructor also has to take departmental considerations into account. The instructor knows that the department wants both good teaching and a specific grade range. While students may form their own expectations of their grades (of which instructors may or may not have a hand in misleading their students and altering expectations), instructors will want to be sure that the *actual* grades that they give for a class meet department standards.

2.3 Departments–The Principals

Departments want the highest possible caliber of student for many reasons. Better quality students are expected to make better quality alumni in the sense that their average expected lifetime salary will be higher than lower quality students. This may lead to higher expected contributions to the endowment. Better quality students will bolster the departments reputation and attract better students in the future. For these reasons, the department may want high average grades.

However, lower average grades are sometimes thought to signify a more difficult program which can also increase the reputation of the department's graduates. This would be especially true of a graduate who stands out because she has high grades compared to a lower departmental average. For example, a graduate who has a 3.8 GPA compared to a departmental average of 2.8 looks much better to prospective employers than a graduate with the same GPA that graduated from a department with a 3.5 departmental average.

To further complicate matters, departments want to be sure that they can both attract and retain students that will do well. The expectation of higher grades might attract many students to the major, but the department only wants to retain those that will be successful in that program. Labor market conditions in different fields will also play a role. (Freeman, 1999) Therefore, each department has their own unique grade distribution that they strive to achieve which depends on current and former student quality and labor market characteristics.

Department chairs and the provost face similar constraints when recruiting and retaining faculty. They want faculty that are both good teachers and good researchers (not to mention hiring these faculty within their budget). They want good teachers that will help produce good students, and they want good researchers that will bolster the reputation of the department as well as bring in external funds. As far as grades and SETs are concerned, the department chairs and provost want their faculty to meet their grading standards while also obtaining high SETs. These can be affected explicitly through department or university mandates or implicitly.

3 Adjustments to SETs and Econometric Issues

Most suggestions for fixing the system involve some sort of adjustment to the final numerical evaluation score. These adjustments are not just based on grades, but other factors as well like difficulty, class size, and composition of the particular class in terms of majors/minors or elective-takers. In fact, the dataset used in this paper already employs such adjustments in their final reports that are sent to department chairs. The relative expected grade adjustment depends on the particular form used and is recalculated each academic year. It typically ranges between 0.25 and 0.45. An example of the adjustment actually used at the University of Washington is presented in Figure 1.[4]

Different institutions choose different methods of adjustment or none at all. For example, the IDEA evaluation system from the University of Kansas adjusts scores for "student motivation to take the class regardless of who taught it, student work habits, class size, student effort not attributable to the instructor, and course difficulty not attributable to the instructor."[5] The Student Instructional Report II (SIR II) from the Educational Testing Service (ETS) provides national comparative data to aid administrations with evaluating results, but offers no direct adjustment.[6] In both of these widely-used datasets, there is no

[4] This figure is reproduced from the OEA and can be found at: `http://web.archive.org/web/20021226003504/www.washington.edu/oea/iasadjst.htm`.

[5] For further details, visit: `http://www.idea.ksu.edu/StudentRatings/index.html`.

[6] See `http://www.ets.org/` for further details.

adjustment or emphasis placed on grades, leaving their adjusted results suspect as well as providing no disincentive for grade inflation.

Most existing numerical adjustments or suggestions for adjustments come from OLS coefficient estimates, which may be biased for several reasons. This has led many studies to attempt to correctly identify the effect of grades on SETs through several different econometric techniques.

One approach is to try to account for the possible endogeneity of relative expected grades as our explanatory variable. In particular, early studies that only use OLS may be providing biased estimates since, "if productivity (an omitted explanatory variable) causes higher evaluations (the dependent variable) and higher grades (an included explanatory variable), then OLS estimates will be biased and inconsistent." (Krautmann and Sander, 1999) Seiver (1983) is the first study to attempt to tackle this issue. After estimating his model using a two-stage least squares (TSLS) approach, he finds that the relationship between expected grades and SETs disappears. Nelson and Lynch (1984) report similar findings. In both of these studies, however, the validity of the instruments used as identifiers have come under question. In particular, the authors use different instructor attributes as supposedly exogenous instruments, when in all likelihood those attributes also affect SETs.

Krautmann and Sander (1999) also use a TSLS approach and find that a one-point increase in average expected GPA has a modest, but significant, impact on evaluations. Using the level of the course as an instrument in their TSLS estimation, this one-point increase yields over half a standard deviation increase in the instructor's SET score; in the OLS estimation, this one-point increase yields almost a full standard deviation increase in the instructor's SET score. After conducting a Hausman test for endogeneity, they conclude that their expected grade variable is not endogenous, and therefore, OLS estimates are appropriate.

However, there are questions about the validity of the level of the course as an instrument as well. There may be selection of instructors into courses, either by the instructors themselves or by their departments. If better instructors choose or are assigned to higher level courses, then the level of the course is not exogenous. If we believe that the course level is not completely exogenous, then the Hausman test is inconclusive, and therefore, the expected grade variable may still be endogenous.

In all cases, it is important to note that most authors believe the hypothetical direction of any bias caused by this potential endogeneity is positive. If there is omitted teacher productivity that causes both higher grades and higher SETs, then the TSLS coefficient on the expected grade variable should be smaller than if the model was estimated using OLS. However, this would only be true if grades accurately reflected learning in the class; this hypothesis is debatable, so the direction of bias is uncertain. (Ewing and Kochin, 2008) In other words, a better instructor's grading policy may be independent of the quality of their students; if better teachers expect better performances from their students knowing that their own quality has an impact, they may adjust their grading standards downward. This may lead to poor instructors overcompensating by inflating grades even more to achieve higher evaluations.

Another possibility would be that students enjoy classes taught by better instructors more than classes taught by poor instructors. This enjoyment of the class may actually take the place of learning rather than complement learning, and thus, students may receive lower grades in classes taught by better instructors. (Langbein, 2007)

The second major approach to identifying the effect of grades on SETs is to use a limited dependent variable (LDV) model to account for the qualitative and discrete nature of SETs when using "questionnaire"-level data. In particular, one must be cautious when interpreting results from a linear model when the dependent variable is discrete, limited, and qualitative since OLS can lead to spurious inferences. For example, OLS will report only one coefficient on each variable even if there is a substantive difference between giving an instructor a "poor" rating instead of just a "very poor" rating and giving an "excellent" rating instead of just a "very good" rating. The first study that mentions this is Mirus (1973), who finds a strong, significant relationship between expected grade and evaluations. DeCanio (1986) addresses this issue by estimating a multinomial logit model and comparing those estimates to standard OLS results. He finds that the expected grade variable is small and insignificant in both specifications. Mehdizadeh (1990) uses loglinear modeling in an attempt to address these issues in a more generalized framework, however, the actual generalizability of his results are questionable since the only classes examined were the ones taught by one instructor, Mehdizadeh himself.

The most promising study conducted in this vein is Mason et al. (1995). They use an ordered probit approach to correct for potential biases introduced in the multinomial logit/probit models when there is a natural ordering to the discrete responses. Using the first generation of the ETS's SIR questionnaire, they find that expected grade in the course and SETs are significantly positively correlated, but they also find that self-reported current GPA from all college courses and SETs are slightly negatively correlated.[7] In other words, it seems that students with higher existing GPAs from their other courses actually give lower evaluations than students with lower existing GPAs. While the effect is small and depends on the particular specification, it does illustrate that a measure of the student's expected grade relative what they are accustomed to receiving may be the appropriate determinant of SETs. It should be noted, however, that none of these limited dependent variable studies takes into account the potential endogeneity of the expected grade variable.[8]

The third and most recent major approach attempts to account for unobserved heterogeneity in the unidentifiable characteristics of each teacher, course, etc. These unobserved characteristics could be attributing to the overall evaluation of the instructor, and by not

[7]The authors tested models with three different measures of the "evaluation" variable: the student's perception of the quality of the lectures, the student's perception of the overall value of the course, and the student's overall rating of the quality of instruction. The current GPA variable was only found to be significant in the first two specifications, with the first specification yielding a positive relationship (but very small) and the second specification yielding a negative relationship (but several times bigger in absolute value than the previous coefficient).

[8]Given that instructors and departments in my dataset only see the mean SET and the mean relative expected grade scores, I do not adopt an LDV approach in this paper. For a discussion of the pros and cons of employing "class"-specific vs. "student"-specific data, see Isely and Singh (2005).

controlling for them, the coefficient on the expected grade variable may be biased. An interesting study along these lines is Hamermesh and Parker (2005). The authors obtain independent measures of the "beauty" of instructors and find that more beautiful instructors do indeed receive higher evaluations, *ceteris paribus*. The authors admit that it is impossible to conclude whether this is a productivity or a discrimination effect.

Isely and Singh (2005) use an instructor fixed effects model and find a significant relationship between expected grade and evaluations. As mentioned earlier, the authors were able to calculate a measure of relative expected grade by taking the average expected grade and subtracting away the average cumulative GPA of each class. As with Mason et al. (1995), the authors find that the relationship between existing GPAs and SETs is actually negative, giving further support for the use of a *relative* expected grade variable. McPherson (2006) also uses a fixed effects model with instructor, course, and semester fixed effects. He also notes that the model estimated in Isely and Singh (2005) does not allow for a comparison across instructors, but rather only allows for a comparison across courses taught by the same instructor.[9]

3.1 Department Effects

The majority of the studies mentioned are confined to just one department (usually Economics). Freeman (1999) shows that "grade divergence" between departments is nothing more than a market outcome in the sense that different departments face different constraints and thus have different incentives for giving the grades that they do. Specifically, Freeman uses National Center for Education Statistics (NCES) data to estimate the impact of salary characteristics on GPA and confirms that graduates from fields of study with higher expected incomes also have lower grades, and vice versa.

Therefore, I believe that "department"-specific effects have an important context when discussing this issue of SETs and relative expected grades, and I am the first, to the best

[9] Both studies do address the issue of endogeneity in their models.

of my knowledge, to include this issue of department effects in the literature encompassing evaluations and grades. For instance, at the University of Washington, grade inflation is perceived as a university-wide trend (Gillmore, 1995; Taggart, 1995), and certain departments have much higher grades than others.[10] This becomes even more important when discussing adjustments to evaluations since the penalties or rewards for high or low grades may be based on department characteristics instead of instructor characteristics if the adjustments are made relative to university-wide grade averages.

In addition to this potentially harmful or beneficial effect to instructors of adjusting evaluations according to university-wide instead of department-wide characteristics, there is a similar underlying effect on students. Students in departments that consistently achieve higher average grades than other departments benefit from the structure of most university-wide honor distinctions. Specifically, within each college at the University of Washington, the top 0.5% of the graduating class receive the honor of graduating summa cum laude, the next 3% graduate magna cum laude, and the next 6.5% graduate cum laude.[11] This means that for a student in the UW College of Arts & Sciences class of 2007 to graduate with any of the above three baccalaureate honors, she would have to have a cumulative GPA of at least 3.75. Therefore, certain departments will have a disproportionately large amount of graduating students with honors if they inflate grades.

[10]For example, at all course levels, average grades in the Dance department are above 3.75. (Gillmore, 1998). This raises a question about whether this is grade inflation or if the quality of the students in the Dance classes is just better than the quality of students in other departments. One argument in favor of better quality Dance students is that the quality of students is readily apparent to all students in those types of classes, and thus, "peer-effects" may exist in the sense that a student who knows he/she is a low-quality dancer will most likely not take a Dance class. Whereas in Mathematics classes, the quality of your peers may be masked. Another argument could be that Dance classes are typically not required for non-Dance majors, whereas some Mathematics classes usually are required for non-Mathematics majors. Whatever the case may be, the important point for this study is to note that there *exists* a difference in grades given between departments that has until now been ignored in suggested SET adjustments.

[11]See `http://www.washington.edu/students/reg/HonorsReqS.html`.

3.2 Data

The majority of the studies mentioned also suffer from very small data sets (the largest being the McPherson (2006) dataset comprising 607 classes taught by 35 instructors over 17 semesters.) The data used in the present study comes from the University of Washington's Office of Educational Assessment. The UW OEA's "Instructional Assessment System" is used in over 12,000 UW classes annually, and at over 50 other post-secondary institutions. In addition, the existing forms have been in use since 1995. Figure 2 shows the questions on the Small Lecture/Discussion evaluation form, the most common form used at the University of Washington.[12] When available, other characteristics outside of the evaluation forms such as average actual grades, time of course, etc. were merged onto the OEA data using data provided by the UW Registrar's office.

For this analysis, I examine data from the entire College of Arts & Sciences at the University of Washington for each quarter from Autumn 1996 through Summer 2006. The College of Arts & Sciences at UW offers classes in 50 departments.[13] For this time frame I have evaluation data for 8,196 unique instructors that teach 58,497 classes.[14] After dropping observations for which the instructor's rank could not be identified or for which there are fewer than three observations per instructor[15], the reduced sample becomes 5,454 instructors teaching 53,658 classes. This larger dataset makes the results more robust for generalization across universities.

[12]For a complete list of all the different types of evaluation forms offered by the Office of Educational Assessment, visit http://www.washington.edu/oea/services/course_eval/index.html.

[13]More information on the College of Arts & Sciences can be found at their website: http://www.artsci.washington.edu/.

[14]The University of Washington requires that faculty have at least one class per year in which he/she teaches be evaluated by their students. Many departments require TAs to have evaluations for every class they teach. I report and discuss results both with and without graduate TAs as one way of determining whether or not there is a selection issue here. I present both given that graduate students are more likely to teach lower-level classes, which is of interest in and of itself.

[15]This is for relevant comparisons to the fixed-effects regressions. Additionally, three observations is the requirement proposed by McPherson (2006).

4 Estimating the Effect of Relative Expected Grade on SETs

The determinants of SETs on the part of the students are intimately related since a student's relative expected grade goes hand in hand with measurements of learning, enjoyment, and effort. So as to avoid problems with collinearity as well as to highlight the importance of relative expected grade on SETs, the only "student"-level variables that are included from the actual SET form are relative expected grade (question #23 in Figure 2) and their answer to whether they are a major, minor, etc. (question #31 in Figure 2) . Additionally, the anonymity of the SETs makes it impossible to match outside student variables to the answers on the questionnaires. To make the measurements consistent, average relative expected grade and average SETs at the "class"-level are used in the estimations.

The characteristics of the class can be used to proxy for student enjoyment and learning and have been linked to course evaluations in other studies.[16] The class-level variables I have available for this study are the level of the class (lower-level classes are defined as 100- or 200-level classes, upper-level classes as 300- or 400-level classes, and graduate-level classes as 500-level classes), the class enrollment, the number of times per week the class meets (once, several times, or daily), and the time the class meets (morning, afternoon, or night). Additionally, I am able to distinguish between full-time classes (defined as credit hours ≥ 3) and seminar classes (defined as credit hours < 3) as well as large-lecture classes with quiz sections and/or labs and regular classes.

The only "instructor"-level variables that are available are the instructor's rank[17] and the actual grade distribution given in the class. However, the ability to follow instructors over time in this panel enables me to use instructor fixed-effects to pick up other time-invariant instructor characteristics.

[16]See Siegfried and Vahaly (1975), Danielsen and White (1976), Marsh et al. (1979), Hamilton (1980), Becker and Powers (2001), and McPherson (2006).

[17]In the estimations, the omitted category is full professor.

Several different versions of SET score were tested as a potential dependent variable from items 1-4 on the evaluation form. The pairwise correlations between each of the items is at least 0.8 for each pair. Given this high correlation and that the OEA combines and reports the median of all four items in addition to adjusting the combined score, the combined score on items 1-4 will be used as the dependent variable in all specifications unless otherwise noted.

4.1 The Basic OLS Model

The basic estimating equation for the OLS model is:

$$SET_i = \beta_0 + \beta_1 G_i^e + \sum_{k=2}^{K} \beta_k x_{ik} + \epsilon_i, \tag{1}$$

where SET_i is the SET score for class i, G_i^e is the average relative expected grade in class i, x_{ik} includes all the other explanatory variables mentioned above, and ϵ_i is assumed to be well-behaved.

4.2 Testing for Endogeneity of Relative Expected Grade

The traditional instruments used for testing for the potential endogeneity of the grade variable in various studies are some measure of the level of the course, actual grades in the class, and/or cumulative GPA (Krautmann and Sander, 1999; McPherson, 2006; Isely and Singh, 2005). For an instrument to be valid, it has to be correlated with the endogenous variable, but uncorrelated with the unobserved variables (in this case, teacher productivity) that are thought to be in the error term. The usage of actual grades that result from a particular class as the instrument for expected grade in that particular class, however, raises many of the same issues that have already been discussed. Namely, actual grades and SETs may be affected by the same unobserved teacher productivity characteristics. Using cumulative GPA may introduce selection issues depending on which students actually filled out the SET questionnaire. The level of the course is also problematic due to the problem of selection into a course by either the instructor or the department; therefore, the level of the course may be correlated with the unobserved teacher productivity.

The instrument I propose is similar to the course effect in McPherson (2006). There are certain effects on relative expected grade that are due to the particular course in question, e.g., holding all else constant, the average relative expected grade in Intermediate Microeconomics may be higher or lower than the average relative expected grade in Intermediate Macroeconomics due to the differences in the material that is covered. In order to use the average course grade as an instrument for relative expected grade in a particular course, however, one must control for any other observations of that course that were taught by the same instructor. Otherwise, the average course grade will be correlated with unobserved teacher productivity. Therefore, the instrument I propose for relative expected grade is the average of the actual grades in other sections of the same course taught by different instructors. Students may use the information on average course grades regardless of instructor in developing their own grade expectations in a particular class, and I exclude any observations that were taught by the instructor in question. Therefore, the average course grade in other sections taught by different instructors satisfies the two requirements for a valid instrument, it is correlated with the suspected endogenous variable and it is uncorrelated with the omitted variable.

Given this particular instrument for relative expected grade, the second-stage regression for the TSLS model is:

$$SET_i = \beta_0 + \beta_1 \widehat{G_i^e} + \sum_{k=2}^{K} \beta_k x_{ik} + \epsilon_i, \tag{2}$$

where $\widehat{G_i^e}$ is obtained from the following first-stage regression:

$$G_i^e = \gamma_0 + \gamma_1 G_i^o + \sum_{k=2}^{K} \gamma_k x_{ik} + \upsilon_i. \tag{3}$$

All notation is as before with the addition of G_i^o, the average grade in all other sections of the same course. Also, υ_i is assumed to be well-behaved.

4.3 Including Fixed Effects

As mentioned above, the estimates do not account for unobserved heterogeneity among instructors or departments. Therefore, a model that includes fixed effects on the part of the instructor, quarter, and department may be warranted. It should be noted that McPherson (2006) includes course fixed effects in his estimations. I chose to not include course fixed effects for several reasons.

First, the level of the course and the department fixed effect is included in my estimations, which encompasses part of the course characteristics. Second, it is the variation in the course characteristics that will partially identify the variation in relative expected grade. In other words, when controlling for instructor fixed effects, the resulting estimated coefficient on relative expected grade can be thought of as an instructor expecting higher evaluations if they teach classes that have higher expected grades. This coefficient can potentially be exploited by an instructor to achieve higher evaluations.

Since the instructor fixed effect is most likely picking up more than just unobserved teacher productivity, e.g., personality, looks, etc., the fixed effects approach can partially be viewed as an alternative approach to the TSLS approach. It is picking up the part of the error term that is constant for all classes an instructor teaches. The coefficient on relative expected grade that remains can be thought of as the effect of a change in grading leniency on evaluations.

Following McPherson (2006), I estimate the following equation:

$$SET_{ijdt} = \beta_0 + \beta_1 G^e_{ijdt} + \alpha_j + \lambda_d + \tau_t + \sum_{k=2}^{K} \beta_k x_{ijdtk} + \epsilon_{ijdt}, \tag{4}$$

where α_j is the fixed effect specific to instructor j, λ_d is the fixed effect specific to department d, and τ_t is the quarter-specific fixed effect. Everything else is as before, except that now we index the observations based on the three types of fixed effects.

5 Results

5.1 Descriptive Statistics

Table 1 gives descriptive statistics for all classes in the sample. Notice that on average, there is effectively no difference between the adjusted and unadjusted evaluation scores, which is by design. The mean adjustment amount is 0.004 with a standard deviation of 0.254. Recall that there are other types of adjustments in addition to relative expected grade, so this should not come as a huge surprise. However, saying that there is no difference between the two is incorrect since the adjusted score is calculated as a linear combination of the evaluation score and the adjustment factors that is constructed so that the averages across the university are the same for a given period. Given that the OEA reports both the instructor's unadjusted and adjusted scores to each department for every class evaluated, however, we can assume that the department chairs and provost look at both figures when deliberating over raises, tenure decisions, etc.

Also, notice that the average relative expected grade is more than one standard deviation above the answer "average" (which would be 4 on the scale for that particular variable.) In other words, for any given class, the answer to the question "Relative to other college courses you have taken: do you expect your grade in this course to be: much higher...much lower?" is most likely "higher". The most likely culprit is overconfidence on the part of the general student.

To test the hypothesis that students in general are overconfident, I first calculated the difference between each class' expected grade and actual grade. Then, I tested

$$H_0 : \text{(Expected Average Grade - Actual Average Grade)} \leq 0 \text{ vs.}$$

$$H_A : \text{(Expected Average Grade - Actual Average Grade)} > 0.$$

The average difference was 0.118 with a standard deviation of 0.25. With such a large number of classes, I was able to conclude that students are indeed overconfident in their expectations. Nowell and Alston (2007) discuss this issue in much more detail, noting that the degree of overconfidence depends on class, instructor, and student characteristics. Rather than estimate a full model of overconfidence as they do in their study, I conducted the same hypothesis tests as above with the population broken up by instructor rank and then again by course level.

Table 2 shows that students' expectations seem to be furthest off the mark when the instructor is either a graduate student or a non-tenure-track lecturer.[18] This could give support to the idea that these instructors may have more incentive to inflate grades to achieve higher evaluations than regular faculty (or at least make it appear that way to their students), but one must also take into account the course effects on overconfidence. Graduate students and lecturers are more likely to teach lower-level courses than upper-level courses, and Table 3 shows that the overconfidence problems are typically relegated to the lower-level courses. Further inspection[19] suggests that this overconfidence stems from both the expectations-forming "inexperience" of the students that take the lower-level courses and that graduate student instructors in particular might be helping inflate grade expectations.

Table 4 breaks out the relevant variables by instructors' rank. At first glance, the hypothesis that lower ranking instructors may be able to influence their SET scores by inflating grades seems plausible. Assistant and Associate Professors are evaluated higher than full Professors and have students that expect higher grades. A similar relationship exists with Lecturers relative to all tenure-track faculty. The relationship of higher grades and higher evaluations seems to break down with graduate students, but this may not necessarily by the case since it is unclear what their evaluation score would be otherwise. Students may be evaluating other aspects of the course (given that quiz sections and labs are taught by

[18]If the number of classes taught by graduate students seems a little high, recall that many departments require *every* class taught by a graduate student be evaluated, whereas that is not the case (at least at the level of the College of Arts & Sciences) for all classes taught by tenure-track faculty.

[19]Results available upon request.

graduate students), or they may be compensating for experience already as they fill out the SET questionnaires. Whatever the case may be, it is clear that controlling for instructors' rank will be important in our estimations.

Figure 3 plots average relative expected grade against average evaluation score, and each point represents a department in the College of Arts & Sciences. Even though there are as of yet no controls for other factors, the positive trend and the large difference between some departments in terms of grades and evaluations in Figure 3 provides a compelling argument for accounting for department effects when estimating the impact of relative expected grade on SETs. For example, of the six departments that have average unadjusted evaluations below 3.8, only one is not underneath the "Natural Science" sub-college umbrella. Likewise, none of the top five departments on the relative expected grade scale are in the natural sciences.

5.2 OLS Results

The results from estimating Equation (1) where Graduate TAs are included are presented in Table 5. The first three specifications in Table 5 are restricted to the subsample of Economics courses taught during this time frame and are included for comparison to Krautmann and Sander (1999), Isely and Singh (2005), and McPherson (2006), respectively.[20] Note that each specification includes relative expected grade, which is measured on a slightly different scale than expected grade, so caution should be taken with a direct comparison of the coefficients between studies. However, the direction and general magnitudes of all the independent variables are similar in scope to those three studies.[21] The results of estimating the full model for just the Economics department is given in Column (4).

[20]There is a potential issue of censoring of the dependent variable given that the mean SET score is close to the upper limit of the possible evaluation range. This "Lake Woebegon Effect" implies that a Tobit model with censoring at 5 may be needed. In my Tobit estimations, less than 1% of the sample is predicted to be right-censored and the coefficients from the Tobit and OLS estimates are very similar. Therefore, I report OLS estimates.

[21]It should be noted that McPherson estimated separate regressions for lower- and upper-level classes. His reasoning comes from a Chow specification test that indicated his data should not be pooled into all levels. I performed a similar test and reached the opposite conclusion with my data.

Column (5) in Table 5 gives the results of the OLS estimation for the entire College of Arts & Sciences. Relative to just the Economics department, there are several striking differences that arise when estimating this model for the entire College. The first big difference is in the relative expected grade coefficient. These estimates are not yet adjusted for potential endogeneity but this could indicate several things, one of which is that for whatever reason, Economics instructors may have a higher incentive to inflate grades. Additionally, the coefficients are in the typical range of the OEA's adjustment. Recall that for a given class, the OEA adjusts instructors' evaluation scores by approximately 0.25 to 0.45 points for every point away from the university-wide average relative expected grade.[22]

The second big difference is the assistant professor coefficient. When isolating the Economics department, the coefficient is significantly negative and relatively large. When looking at the entire college, the coefficient is significantly positive but small. This could be due to the fact that during this time frame, there are only 13 assistant professors in the Economics department and without instructor fixed effects, unobserved individual teacher characteristics may be affecting this coefficient. The difference in directions may indicate that Economics students may value experience more than students in the rest of the college, but more likely, the large difference will be washed out when including instructor fixed effects.

Columns (6), (7), (8), and (9) break the College of Arts & Sciences down by department into sub-colleges: Arts, Social Sciences, Natural Sciences, and Humanities. The differences in the coefficients on relative expected grade suggest that department effects do indeed play a role in determining the final effect of relative expected grade on SETs. This could be done for all 50 departments as was done for Economics, but several departments have much fewer observations over a given period of time than others, e.g., the American Indian Studies department has only 174 observations over the ten year time frame.

Table 6 gives the results of the OLS estimations when Graduate TAs are not included.

[22] See Figure 1.

Most of the coefficients remain similar in magnitude and significance. The largest changes as far as the relative expected grade coefficient is concerned occurs in the Arts and Humanities sub-colleges.

5.3 TSLS Results

The first-stage regression results from estimating Equation (3) are reported in Table 7. In all of the different subsamples, the average grade in other sections of the same course is a highly significant determinant of relative expected grade. Additionally, it does not appear that this is a “weak” instrument when using the rule-of-thumb that the F-statistics should be above 10 in each subsample. (Staiger and Stock, 1997)

The second-stage regression results from estimating Equation (2) are reported in Table 8. Hausman tests were conducted for all subsamples, with the null hypothesis being that relative expected grade is exogenous versus the alternative hypothesis that relative expected grade is endogenous. At the 5%-level, this null hypothesis is rejected for all subsamples except for the Economics department. At the 10%-level, it is rejected for the Economics department as well. Therefore, if this instrument is accepted as a valid instrument, then the Hausman tests imply that relative expected grade is indeed endogenous and that the TSLS estimates, and not the OLS estimates, should be used.

Table 8 shows that the coefficients on relative expected grade are highly significant and roughly 0.2 to 0.4 points higher than the OLS estimates for each subsample. If there is no measurement error on the part of the relative expected grade variable, this implies that instructors actually have *more* incentive to inflate grades in order to achieve higher SET scores than previously thought.

Table 9 provides an compelling case for perhaps adopting a fixed effects approach in place of TSLS. At the very least, while the instrument used in this study theoretically meets the requirements of a valid instrument, it is heavily dependent on having multiple observations of the same course taught by different instructors. Excluding graduate TAs, at least in the case of the Economics department, excludes many of the observations that strengthened this particular instrument. There are typically many more sections offered of lower-level courses than upper-level courses over a given period of time, and these lower-level courses are precisely the ones that graduate students teach most often.

Table 10 shows that for the specifications where we may be willing to accept the instrument as valid, lecturers and tenure-track professors may have a large incentive to inflate grades even after accounting for endogenous teacher productivity.

5.4 Fixed Effects Results

Fixed effects was chosen over random effects based on the results from Hausman specification tests, a small difference between the coefficients on relative expected grade between the two specifications, and since the fixed effects results for a particular instructor in a particular department may be useful in the SET adjustment process.

The results from estimating Equation (4) are presented in Table 11. Table 11 shows that the coefficient on relative expected grade decreases by about a tenth of a point from the OLS estimates. This should come as no surprise, since the constant instructor and department characteristics that were initially unobserved are now being picked up in the fixed effects. However, the fact that it remains positive and significant suggests that instructors do indeed have an incentive to inflate grades to achieve higher evaluations.

The coefficient on relative expected grade can be viewed as a lower bound to the range of the true impact of grading leniency on evaluations in this fixed effects model. While the TSLS estimates are closely linked to the strength of the instrument for relative expected grade, the fixed effects estimates are not dependent upon discovering a strong, valid instrument. Fixed effects also takes into account much more than unobserved teacher productivity, giving a clearer picture of the impact of grade inflation on evaluations.

However, having the range of estimates from both fixed effects and TSLS is useful since an implicit assumption in the fixed effects model is that teacher productivity is constant over all classes taught by a particular instructor. Since it is plausible that instructors may be better at teaching some courses than others, teacher productivity may not be completely controlled for with instructor fixed effects. Likewise, instructor fixed effects holds teacher productivity constant over time, which is most likely not the case.

Having department fixed effects may give a good idea about the ability of a particular instructor within a particular department to affect his/her evaluations through more lenient grading. However, when examining the results, one has to be careful not to attach the wrong meaning to the department fixed effect. It is not necessarily true or false that if instructor A from department X and instructor B from department Y have similar instructor fixed effects but much different department fixed effects, that we can conclude that the instructor in the department with the higher department fixed effect is a better quality instructor. While the department fixed effect may pick up some of the average quality of its instructor pool, it is also picking up other attributes that have nothing to do with instructor quality but nevertheless influence SETs.

Column (2) of Table 11 shows that when excluding graduate TAs, there is little change to the relevant coefficients. This is not surprising since the instructor fixed effect accounts partially for the experience of graduate students.

The fixed effects results suggest that a starting point for adjustment may be the 0.240 coefficient on relative expected grade, with additional adjustments based on department and quarter fixed effects. Additionally, instructor fixed effects can be used to isolate those instructors who seem to have the highest levels of constant teacher productivity. (McPherson, 2006)

6 Conclusion

It seems that no matter which estimation procedure one chooses, there is a significant positive effect of relative expected grades on evaluation scores. The magnitude of this impact ranges from a 0.240 point increase in SET score for every point increase on the relative expected grade scale to a 1.123 point increase for a particular sub-college after excluding graduate TAs.

This positive impact is the strongest in the TSLS estimations after controlling for the potential endogeneity of the relative expected grade variable due to unobserved teacher productivity, if the instrument of average course grade in all sections of the same course not taught by that instructor is accepted as a valid instrument. This raises a question about what students actually value in their outcomes from a particular course, learning or enjoyment.

This positive effect even remains after including department and instructor fixed effects. The unobserved characteristics of the department and the instructor are important determinants of SETs. This implies that it is necessary to account for instructor and department characteristics if attempting to make cross-instructor and cross-department comparisons with SETs.

Current adjustments to evaluations for possible grade inflation at most universities, if used at all, are not typically done on a departmental basis but rather on a university-wide basis. The range of estimates for grading leniency's impact on evaluations will depend on the estimation procedure used, but it is clear that OLS may not be sufficient. My results suggest that there is a need to update this procedure to at least make note of department differences if not directly adjust for them. Otherwise, the adjustment mechanism may be adding more noise to an already noisy evaluation instrument.

References

Becker, William E. and John R. Powers (2001). "Student Performance, Attrition, and Class Size Given Missing Student Data." *Economics of Education Review*, 20:377–388.

Danielsen, Albert L. and Rudolph A. White (1976). "Some Evidence on the Variables Associated with Student Evaluations of Teachers." *The Journal of Economic Education*, 7(2):117–119.

DeCanio, Stephen J. (1986). "Student Evaluations of Teaching: A Multinomial Logit Approach." *The Journal of Economic Education*, 17(3):165–176.

Ewing, Andrew M. and Levis Kochin (2008). "Learning, Grades, and Student Evaluations of Teaching in an Economics Course Sequence."

Freeman, Donald G. (1999). "Grade Divergence as a Market Outcome." *The Journal of Economic Education*, 30(4):344–351.

Garicano, Luis (2008). "Hierarchy." In Steven N. Durlauf and Lawrence E. Blume, eds., "The New Palgrave Dictionary of Economics," Palgrave Macmillan, 2nd edn.

Gillmore, Gerald M. (1995). "Grades." Tech. Rep. 95-N3, University of Washington Office of Educational Assessment.

——— (1998). "Average Grades." Tech. Rep. 98-N1, University of Washington Office of Educational Assessment.

Hamermesh, Daniel S. and Amy Parker (2005). "Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity." *Economics of Education Review*, 24:369–376.

Hamilton, Lawrence C. (1980). "Grades, Class Size, and Faculty Status Predict Teaching Evaluations." *Teaching Sociology*, 8(1):47–62.

Hostetler, Karl D., R. McLaran Sawyer, and Keith W. Prichard, eds. (2004). *The Art and Politics of College Teaching: A Practical Guide for the Beginning Professor*. Peter Lang Publishing, Inc., 2nd edn.

Isely, Paul and Harinder Singh (2005). "Do Higher Grades Lead to Favorable Student Evaluations?" *The Journal of Economic Education*, 36(1):29–42.

Krautmann, Anthony C. and William Sander (1999). "Grades and Student Evaluations of Teachers." *Economics of Education Review*, 18(1):59–63.

Langbein, Laura I. (2007). "Management by Results: Student Evaluation of Faculty Teaching and the Mis-measurement of Performance."

Lazear, Edward (1995). *Personnel Economics*. The MIT Press.

Marlin, James W., Jr. (1987). "Student Perceptions of End-of-Course Evaluations." *The Journal of Higher Education*, 58(6):704–716.

Marsh, Herbert W., Jesse U. Overall, and Steven P. Kesler (1979). "Class Size, Students' Evaluations, and Instructional Effectiveness." *American Educational Research Journal*, 16(1):57–70.

Mason, Paul M., Jeffrey W. Steagall, and Michael M. Fabritius (1995). "Student Evaluations of Faculty: A New Procedure for Using Aggregate Measures of Performance." *Economics of Education Review*, 14(4):403–416.

McPherson, Michael A. (2006). "Determinants of How Students Evaluate Teachers." *The Journal of Economic Education*, 37(1):3–20.

Mehdizadeh, Mostafa (1990). "Loglinear Models and Student Course Evaluations." *The Journal of Economic Education*, 21(1):7–21.

Mirus, Rolf (1973). "Some Implications of Student Evaluation of Teachers." *The Journal of Economic Education*, 5(1):35–37.

Moore, Melanie and Richard Trahan (1998). "Tenure Status and Grading Practices." *Sociological Perspectives*, 41(4):775–781.

Nelson, Jon P. and Kathleen A. Lynch (1984). "Grade Inflation, Real Income, Simultaneity, and Teaching Evaluations." *The Journal of Economic Education*, 15(1):21–37.

Nowell, Clifford and Richard M. Alston (2007). "I Thought I Got an A! Overconfidence Across the Economics Curriculum." *The Journal of Economic Education*, 38(2):131–142.

Seiver, Daniel A. (1983). "Evaluations and Grades: A Simultaneous Framework." *The Journal of Economic Education*, 14(3):32–38.

Siegfried, John J. and John Vahaly, Jr. (1975). "Sample Bias of Unannounced Student Evaluations of Teaching." *The Journal of Economic Education*, 6(2):137–139.

Staiger, Douglas and James H. Stock (1997). "Instrumental Variables Regressions with Weak Instruments." *Econometrica*, 65(3):557–586.

Taggart, Thomas (1995). "Faculty Views of the Grading System and "Grade Inflation" at the University of Washington." Tech. Rep. 95-4, University of Washington Office of Educational Assessment.

A Tables

Table 1: Descriptive Statistics

Variable	Mean	Standard Deviation
Unadjusted Evaluation Score	4.013	0.623
Adjusted Evaluation Score	4.012	0.559
Average Relative Expected Grade	4.799	0.653
Average Expected Grade	3.402	0.245
Actual Average Grade	3.283	0.347
Class Size	32.871	44.033
Professor	11.4%	—
Associate Professor	6.7%	—
Assistant Professor	8.0%	—
Lecturer	13.9%	—
Graduate Student	57.7%	—
Lower Level Class	58.8%	—
Upper Level Class	33.5%	—
Graduate Class	7.7%	—
Large Lecture Class	7.2%	—
Quiz Section or Lab	31.8%	—

Table 2: Students' Overconfidence by Instructor's Rank

Rank	Average Difference between Expected and Actual Grades	Standard Deviation	Number of Classes
Professor	0.043	0.249	5,891
Associate Prof.	0.034	0.237	3,455
Assistant Prof.	0.038	0.244	4,048
Lecturer	0.100	0.240	6,823
Graduate Student	0.166	0.253	31,589

Table 3: Students' Overconfidence by Course Level

Course Level	Average Difference between Expected and Actual Grades	Standard Deviation	Number of Classes
One-Hundred Level	0.189	0.247	17,516
Two-Hundred Level	0.156	0.240	14,502
Three-Hundred Level	0.118	0.252	10,358
Four-Hundred Level	0.022	0.240	7,249
Graduate Class	-0.116	0.171	3,326

Table 4: Descriptive Statistics by Instructors' Rank†

Rank	Relative Expected Grade	Evaluation Score
Professor	4.615	4.065
	(0.668)	(0.663)
Associate Prof.	4.733	4.180
	(0.669)	(0.613)
Assistant Prof.	4.740	4.171
	(0.638)	(0.594)
Lecturer	4.940	4.223
	(0.647)	(0.564)
Graduate Student	4.816	3.909
	(0.642)	(0.611)
Total	4.798	4.012
	(0.653)	(0.624)

†These numbers represent the average score by rank across the entire College of Arts & Sciences, with standard deviations in parentheses.

Table 5: OLS Regressions–Including Graduate TAs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Relative Expected Grade	0.452***	0.459***	0.463***	0.465***	0.325***	0.263***	0.298***	0.344***	0.266***
	(0.021)	(0.021)	(0.020)	(0.021)	(0.004)	(0.013)	(0.008)	(0.007)	(0.007)
Associate Professor	0.033	-0.006	-0.020	0.012	0.070***	-0.028	0.044**	0.021	0.053**
	(0.071)	(0.070)	(0.070)	(0.070)	(0.012)	(0.035)	(0.022)	(0.020)	(0.024)
Assistant Professor	-0.215***	-0.216***	-0.211***	-0.222***	0.065***	-0.058*	0.003	0.064***	-0.011
	(0.076)	(0.075)	(0.074)	(0.075)	(0.011)	(0.031)	(0.021)	(0.020)	(0.022)
Lecturer	0.077	0.061	0.036	0.104	0.106***	0.021	0.030	0.170***	-0.005
	(0.077)	(0.076)	(0.074)	(0.078)	(0.010)	(0.033)	(0.022)	(0.017)	(0.020)
Graduate Student	-0.066	-0.000	-0.062	-0.034	-0.113***	-0.017	-0.159***	-0.088***	-0.209***
	(0.054)	(0.054)	(0.048)	(0.054)	(0.010)	(0.032)	(0.019)	(0.017)	(0.020)
Enrollment	-0.002***	-0.001***	-0.001***	-0.000	-0.001***	-0.002***	-0.000***	-0.000***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Response Rate		0.008***	0.009***	0.008***	0.003***	0.001	0.004***	0.004***	0.001***
		(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Percent Major		0.002**	0.001**	0.002**	0.001***	0.001**	0.001***	0.001***	0.002***
		(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Class Meets Once a Week		0.169***	0.182***	0.074	-0.248***	-0.011	-0.083***	-0.087***	-0.354***
		(0.048)	(0.047)	(0.058)	(0.009)	(0.044)	(0.018)	(0.026)	(0.023)
Class Meets Several Times a Week		0.139***	0.122***	0.085**	-0.212***	-0.075*	-0.051***	-0.062**	-0.266***
		(0.035)	(0.033)	(0.041)	(0.008)	(0.041)	(0.017)	(0.025)	(0.010)
Afternoon Class		0.005		0.009	0.020***	0.014	0.047***	-0.003	0.032***
		(0.029)		(0.029)	(0.005)	(0.018)	(0.010)	(0.009)	(0.009)
Evening Class		-0.073		-0.023	-0.023	-0.257***	-0.052*	0.004	0.154***
		(0.060)		(0.063)	(0.014)	(0.078)	(0.028)	(0.021)	(0.031)
Quiz Section or Lab				0.101***	-0.029***	-0.224***	-0.094***	0.056***	0.097***
				(0.035)	(0.007)	(0.025)	(0.013)	(0.012)	(0.016)
Large Lecture Course				-0.284***	-0.119***	-0.090*	-0.135***	-0.081***	0.003
				(0.103)	(0.011)	(0.049)	(0.024)	(0.016)	(0.039)
Upper-Level Class	0.050	-0.022		0.017	0.103***	0.085***	0.085***	0.122***	0.071***
	(0.036)	(0.053)		(0.055)	(0.007)	(0.023)	(0.013)	(0.012)	(0.014)
Graduate Level Class	0.299***	0.160***		0.225***	0.258***	0.199***	0.223***	0.260***	0.320***
	(0.049)	(0.058)		(0.060)	(0.012)	(0.040)	(0.024)	(0.019)	(0.025)
Constant	1.773***	1.005***	1.043***	0.982***	2.390***	2.973***	2.400***	1.985***	3.010***
	(0.117)	(0.135)	(0.134)	(0.134)	(0.024)	(0.095)	(0.048)	(0.046)	(0.044)
Departments Included	Economics	Economics	Economics	Economics	All	Arts	Social Sciences	Natural Sciences	Humanities
Observations	2010	1988	2010	1988	50056	3432	15433	17248	13943
R-squared	0.237	0.288	0.279	0.294	0.237	0.197	0.213	0.203	0.210

The dependent variable in each specification is the average of the unadjusted medians from the first four questions in Figure 2.
Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 6: OLS Regressions–Excluding Graduate TAs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Relative Expected Grade	0.415***	0.408***	0.434***	0.393***	0.313***	0.204***	0.272***	0.406***	0.206***
	(0.048)	(0.048)	(0.047)	(0.047)	(0.006)	(0.018)	(0.012)	(0.011)	(0.010)
Associate Professor	0.025	0.005	0.004	0.026	0.067***	-0.024	0.040**	0.011	0.047**
	(0.070)	(0.070)	(0.070)	(0.068)	(0.011)	(0.035)	(0.020)	(0.020)	(0.023)
Assistant Professor	-0.251***	-0.244***	-0.232***	-0.264***	0.062***	-0.052*	0.006	0.055***	-0.023
	(0.076)	(0.076)	(0.076)	(0.075)	(0.011)	(0.031)	(0.020)	(0.020)	(0.021)
Lecturer	0.022	0.062	0.008	0.116	0.107***	0.036	0.013	0.140***	0.009
	(0.080)	(0.080)	(0.075)	(0.079)	(0.010)	(0.034)	(0.021)	(0.017)	(0.020)
Enrollment	-0.001	-0.000	-0.001***	0.000	-0.001***	-0.003***	-0.000	-0.000***	-0.002***
	(0.001)	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Response Rate		0.006***	0.007***	0.005***	0.004***	-0.000	0.005***	0.005***	0.002***
		(0.002)	(0.002)	(0.002)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Percent Major		0.000	-0.001	0.000	0.001***	-0.000	-0.000	0.001***	0.001**
		(0.002)	(0.001)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Class Meets Once a Week		0.399*	0.458**	0.386*	-0.069***	0.071	-0.036	-0.068*	-0.080**
		(0.230)	(0.223)	(0.224)	(0.018)	(0.066)	(0.032)	(0.038)	(0.035)
Class Meets Several Times a Week		0.080	0.094	0.094	-0.137***	-0.059	-0.081***	-0.114***	-0.100***
		(0.143)	(0.140)	(0.139)	(0.012)	(0.056)	(0.025)	(0.031)	(0.017)
Afternoon Class		-0.082		-0.082	0.031***	-0.015	0.035**	0.007	0.021
		(0.064)		(0.062)	(0.008)	(0.025)	(0.016)	(0.014)	(0.014)
Evening Class		-0.238		-0.142	-0.069***	-0.345***	-0.002	-0.142***	0.113**
		(0.252)		(0.247)	(0.022)	(0.088)	(0.038)	(0.037)	(0.047)
Large Lecture Course				-0.520***	-0.158***	-0.108*	-0.158***	-0.084***	-0.156***
				(0.115)	(0.012)	(0.060)	(0.026)	(0.017)	(0.044)
Upper-Level Class	0.152	0.226		-0.027	0.113***	0.105***	0.122***	0.109***	0.079***
	(0.196)	(0.220)		(0.221)	(0.011)	(0.030)	(0.024)	(0.020)	(0.019)
Graduate Level Class	0.383*	0.437*		0.224	0.215***	0.159***	0.187***	0.237***	0.249***
	(0.220)	(0.236)		(0.234)	(0.015)	(0.046)	(0.031)	(0.024)	(0.028)
Constant	1.835***	1.310***	1.478***	1.673***	2.316***	3.400***	2.502***	1.673***	3.198***
	(0.329)	(0.389)	(0.337)	(0.388)	(0.036)	(0.125)	(0.073)	(0.065)	(0.063)
Departments Included	Economics	Economics	Economics	Economics	All	Arts	Social Sciences	Natural Sciences	Humanities
Observations	370	370	370	370	18992	1771	4971	6932	5318
R-squared	0.299	0.329	0.318	0.366	0.258	0.170	0.192	0.281	0.143

The dependent variable in each specification is the average of the unadjusted medians from the first four questions in Figure 2.
Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 7: TSLS Regressions–First-Stage Results–Including Graduate TAs

	(1)	(2)	(3)	(4)	(5)	(6)
Avg Grade in Other Sections	0.853***	0.742***	1.377***	0.407***	0.524***	0.376***
of Same Course	(0.156)	(0.013)	(0.057)	(0.030)	(0.020)	(0.038)
Associate Professor	-0.325***	0.081***	0.002	0.003	0.068***	-0.015
	(0.082)	(0.015)	(0.052)	(0.026)	(0.024)	(0.035)
Assistant Professor	-0.239***	0.103***	-0.023	0.071***	0.082***	-0.001
	(0.092)	(0.014)	(0.045)	(0.025)	(0.022)	(0.031)
Lecturer	-0.123	0.270***	0.038	0.218***	0.341***	0.154***
	(0.088)	(0.013)	(0.047)	(0.025)	(0.019)	(0.028)
Graduate Student	-0.053	0.258***	0.095**	0.153***	0.195***	0.136***
	(0.063)	(0.012)	(0.045)	(0.023)	(0.019)	(0.028)
Enrollment	0.000	0.001***	0.001**	0.001***	0.000***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Response Rate	0.001	0.001***	0.002**	0.001**	0.001**	0.000
	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Percent Major	0.003***	-0.002***	-0.004***	0.001***	-0.002***	-0.003***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Class Meets Once a	-0.237***	-0.021**	-0.259***	-0.109***	0.011	0.106***
Week	(0.062)	(0.011)	(0.057)	(0.020)	(0.029)	(0.030)
Class Meets Several	-0.148***	-0.039***	-0.242***	-0.112***	-0.119***	0.053***
Times a Week	(0.044)	(0.009)	(0.052)	(0.018)	(0.028)	(0.012)
Afternoon Class	-0.007	-0.009	-0.005	-0.012	-0.011	0.019
	(0.032)	(0.006)	(0.024)	(0.011)	(0.010)	(0.012)
Evening Class	-0.004	-0.041**	0.245**	0.067**	-0.051**	0.003
	(0.069)	(0.017)	(0.101)	(0.031)	(0.025)	(0.040)
Upper-Level Class	-0.253***	-0.038***	-0.154***	-0.074***	0.103***	-0.107***
	(0.061)	(0.008)	(0.030)	(0.014)	(0.014)	(0.018)
Graduate Level Class	-0.336***	-0.480***	-0.688***	-0.305***	-0.222***	-0.460***
	(0.102)	(0.017)	(0.057)	(0.032)	(0.026)	(0.039)
Quiz Section or Lab	-0.193***	-0.039***	0.144***	-0.090***	0.094***	-0.039*
	(0.038)	(0.008)	(0.033)	(0.014)	(0.013)	(0.021)
Large Lecture Course	-0.069	-0.054***	0.047	-0.110***	0.034*	-0.129***
	(0.112)	(0.013)	(0.061)	(0.026)	(0.017)	(0.050)
Constant	2.223***	2.250***	0.623***	3.413***	2.842***	3.594***
	(0.496)	(0.046)	(0.220)	(0.104)	(0.072)	(0.135)
Departments Included	Economics	All	Arts	Social Sciences	Natural Sciences	Humanities
Observations	1926	45687	2979	14114	15485	13109
R-squared	0.096	0.120	0.270	0.043	0.105	0.054
F-Statistic	12.72	389.99	68.53	39.53	113.64	47.11

The dependent variable in each specification is the average of the unadjusted medians from the first four questions in Figure 2. Standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 8: TSLS Regressions–Second-Stage Results–Including Graduate TAs

	(1)	(2)	(3)	(4)	(5)	(6)
Relative Expected Grade	0.701***	0.670***	0.548***	0.650***	0.634***	0.471***
	(0.174)	(0.017)	(0.037)	(0.077)	(0.038)	(0.083)
Associate Professor	0.154	0.051***	-0.009	0.071***	-0.006	0.065**
	(0.096)	(0.015)	(0.046)	(0.028)	(0.024)	(0.028)
Assistant Professor	-0.170*	0.031**	0.001	-0.015	0.054**	-0.012
	(0.096)	(0.013)	(0.040)	(0.027)	(0.022)	(0.025)
Lecturer	0.142	0.020	0.059	-0.062**	0.071***	-0.028
	(0.087)	(0.013)	(0.041)	(0.031)	(0.023)	(0.026)
Graduate Student	-0.015	-0.156***	0.033	-0.179***	-0.078***	-0.226***
	(0.062)	(0.012)	(0.040)	(0.024)	(0.019)	(0.025)
Enrollment	-0.000	-0.001***	-0.002***	-0.001***	-0.000***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Response Rate	0.007***	0.003***	0.000	0.004***	0.003***	0.001***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Percent Major	0.001	0.002***	0.002***	0.001***	0.002***	0.003***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Class Meets Once a Week	0.132*	-0.229***	0.071	-0.032	-0.072**	-0.405***
	(0.074)	(0.010)	(0.051)	(0.021)	(0.029)	(0.027)
Class Meets Several Times a Week	0.119**	-0.195***	-0.016	0.008	-0.008	-0.292***
	(0.050)	(0.009)	(0.046)	(0.020)	(0.028)	(0.011)
Afternoon Class	0.013	0.026***	-0.009	0.049***	0.001	0.029***
	(0.031)	(0.006)	(0.022)	(0.011)	(0.010)	(0.010)
Evening Class	-0.014	0.017	-0.254***	-0.110***	0.009	0.180***
	(0.065)	(0.017)	(0.090)	(0.033)	(0.024)	(0.033)
Upper-Level Class	0.057	0.070***	0.090***	0.096***	0.046***	0.068***
	(0.076)	(0.008)	(0.027)	(0.015)	(0.015)	(0.016)
Graduate Level Class	0.198***	0.297***	0.311***	0.227***	0.228***	0.377***
	(0.067)	(0.015)	(0.052)	(0.033)	(0.023)	(0.038)
Quiz Section or Lab	0.150***	0.014*	-0.187***	-0.058***	0.010	0.124***
	(0.050)	(0.008)	(0.028)	(0.017)	(0.013)	(0.018)
Large Lecture Course	-0.271**	-0.067***	-0.055	-0.096***	-0.083***	0.053
	(0.109)	(0.013)	(0.054)	(0.029)	(0.017)	(0.042)
Constant	-0.195	0.747***	1.326***	0.705*	0.641***	2.022***
	(0.858)	(0.079)	(0.217)	(0.367)	(0.172)	(0.406)
Departments Included	Economics	All	Arts	Social Sciences	Natural Sciences	Humanities
Observations	1926	45687	2979	14114	15485	13109
R-squared	0.250	0.119	0.104	0.104	0.124	0.164

The dependent variable in each specification is the average of the unadjusted medians from the first four questions in Figure 2. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 9: TSLS Regressions–First-Stage Results–Excluding Graduate TAs

	(1)	(2)	(3)	(4)	(5)	(6)
Avg Grade in Other Sections	-0.273	0.693***	1.214***	0.401***	0.638***	0.314***
of Same Course	(0.238)	(0.022)	(0.089)	(0.049)	(0.037)	(0.061)
Associate Professor	-0.287***	0.080***	0.003	-0.013	0.065***	-0.016
	(0.083)	(0.015)	(0.053)	(0.028)	(0.023)	(0.035)
Assistant Professor	-0.251***	0.107***	-0.018	0.060**	0.087***	0.001
	(0.094)	(0.014)	(0.046)	(0.027)	(0.022)	(0.031)
Lecturer	-0.049	0.276***	0.090*	0.231***	0.337***	0.148***
	(0.094)	(0.013)	(0.049)	(0.026)	(0.019)	(0.029)
Enrollment	0.000	0.001***	0.001**	0.001***	0.001***	0.002***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Response Rate	-0.003	0.000	0.001	-0.000	0.000	0.000
	(0.002)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)
Percent Major	0.005**	-0.000*	-0.001	0.003***	-0.001***	-0.002***
	(0.002)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Class Meets Once a	0.100	-0.073***	-0.229**	-0.229***	-0.046	0.047
Week	(0.269)	(0.024)	(0.092)	(0.045)	(0.044)	(0.053)
Class Meets Several	-0.177	-0.065***	-0.122	-0.101***	-0.086**	0.022
Times a Week	(0.156)	(0.015)	(0.077)	(0.033)	(0.033)	(0.024)
Afternoon Class	-0.212***	0.030***	-0.041	-0.012	0.058***	0.041**
	(0.076)	(0.010)	(0.036)	(0.021)	(0.016)	(0.020)
Evening Class	-0.443	0.137***	0.376***	0.138***	0.208***	0.010
	(0.317)	(0.028)	(0.119)	(0.051)	(0.040)	(0.065)
Upper-Level Class	-0.241	-0.046***	-0.118***	-0.024	0.095***	-0.084***
	(0.255)	(0.014)	(0.041)	(0.031)	(0.023)	(0.027)
Graduate Level Class	0.527*	-0.456***	-0.533***	-0.177***	-0.265***	-0.414***
	(0.304)	(0.021)	(0.069)	(0.046)	(0.035)	(0.047)
Large Lecture Course	-0.156	-0.062***	-0.031	-0.086***	0.009	-0.097
	(0.132)	(0.015)	(0.078)	(0.032)	(0.018)	(0.060)
Constant	6.007***	2.373***	0.901***	3.333***	2.390***	3.793***
	(0.810)	(0.076)	(0.335)	(0.170)	(0.122)	(0.211)
Departments Included	Economics	All	Arts	Social Sciences	Natural Sciences	Humanities
Observations	309	15468	1341	3731	5852	4544
R-squared	0.149	0.137	0.171	0.077	0.141	0.078
F-Statistic	3.66	174.74	19.58	22.20	68.57	27.36

The dependent variable in each specification is the average of the unadjusted medians from the first four questions in Figure 2. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 10: TSLS Regressions–Second-Stage Results–Excluding Graduate TAs

	(1)	(2)	(3)	(4)	(5)	(6)
Relative Expected Grade	-0.100	0.927***	0.477***	0.907***	0.959***	1.123***
	(0.907)	(0.034)	(0.064)	(0.130)	(0.063)	(0.231)
Associate Professor	-0.060	0.021	-0.004	0.076**	-0.035	0.063
	(0.280)	(0.017)	(0.046)	(0.030)	(0.026)	(0.042)
Assistant Professor	-0.393	-0.010	-0.001	-0.027	0.015	-0.027
	(0.255)	(0.016)	(0.040)	(0.029)	(0.025)	(0.037)
Lecturer	0.091	-0.066***	0.044	-0.151***	-0.052*	-0.114**
	(0.110)	(0.018)	(0.043)	(0.041)	(0.031)	(0.047)
Enrollment	0.000	-0.001***	-0.003***	-0.000**	-0.000***	-0.004***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Response Rate	0.004	0.004***	-0.001	0.005***	0.005***	0.002**
	(0.003)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)
Percent Major	0.005	0.001***	-0.000	-0.002***	0.002***	0.003***
	(0.005)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)
Class Meets Once a	0.427	-0.069***	0.190**	0.154***	-0.102**	-0.142**
Week	(0.289)	(0.025)	(0.081)	(0.050)	(0.049)	(0.064)
Class Meets Several	0.003	-0.110***	-0.000	0.014	-0.049	-0.151***
Times a Week	(0.219)	(0.016)	(0.067)	(0.035)	(0.038)	(0.029)
Afternoon Class	-0.170	-0.006	-0.025	0.029	-0.037**	-0.018
	(0.211)	(0.011)	(0.031)	(0.023)	(0.018)	(0.026)
Evening Class	-0.269	-0.116***	-0.384***	-0.127**	-0.228***	0.160**
	(0.514)	(0.030)	(0.107)	(0.057)	(0.046)	(0.078)
Upper-Level Class	-0.284	0.086***	0.067*	0.119***	-0.023	0.115***
	(0.344)	(0.015)	(0.036)	(0.032)	(0.027)	(0.034)
Graduate Level Class	0.352	0.323***	0.197***	0.146***	0.183***	0.509***
	(0.434)	(0.022)	(0.061)	(0.046)	(0.032)	(0.083)
Large Lecture Course	-0.603***	-0.060***	-0.062	-0.089**	-0.066***	-0.047
	(0.183)	(0.017)	(0.068)	(0.038)	(0.020)	(0.074)
Constant	4.140	-0.525***	1.944***	-0.473	-0.778***	-1.240
	(4.703)	(0.159)	(0.346)	(0.606)	(0.281)	(1.119)
Departments Included	Economics	All	Arts	Social Sciences	Natural Sciences	Humanities
Observations	309	15468	1341	3731	5852	4544
R-squared	0.165	.	0.101	.	0.053	.

The dependent variable in each specification is the average of the unadjusted medians from the first four questions in Figure 2. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 11: Fixed Effects Regression

	(1)	(2)
Relative Expected Grade	0.240***	0.245***
	(0.004)	(0.006)
Associate Professor	-0.021	-0.024
	(0.019)	(0.019)
Assistant Professor	-0.056**	-0.052**
	(0.022)	(0.023)
Lecturer	-0.009	-0.011
	(0.028)	(0.031)
Graduate Student	-0.062**	
	(0.030)	
Enrollment	-0.002***	-0.001***
	(0.000)	(0.000)
Response Rate	0.001***	0.002***
	(0.000)	(0.000)
Percent Major	0.001***	0.001***
	(0.000)	(0.000)
Class Meets Once a Week	-0.098***	-0.025
	(0.011)	(0.019)
Class Meets Several Times a Week	-0.074***	-0.044***
	(0.009)	(0.013)
Afternoon Class	0.033***	0.015**
	(0.005)	(0.007)
Evening Class	0.072***	0.038*
	(0.013)	(0.022)
Quiz Section or Lab	-0.008	
	(0.008)	
Large Lecture Course	-0.109***	-0.126***
	(0.011)	(0.012)
Upper-Level Class	0.073***	0.105***
	(0.007)	(0.011)
Graduate Level Class	0.209***	0.211***
	(0.012)	(0.015)
Constant	2.907***	3.089***
	(0.065)	(0.094)
Observations	50056	18992
Number of Unique Instructors	5362	1999
Sample Restrictions	Includes Graduate TAs	Excludes Graduate TAs
R-squared	0.160	0.210

The dependent variable in each specification is the average of the unadjusted medians from the first four questions in Figure 2. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Instructor, Department, and Quarter fixed effects are included, but not reported. Results available upon request.

B Figures

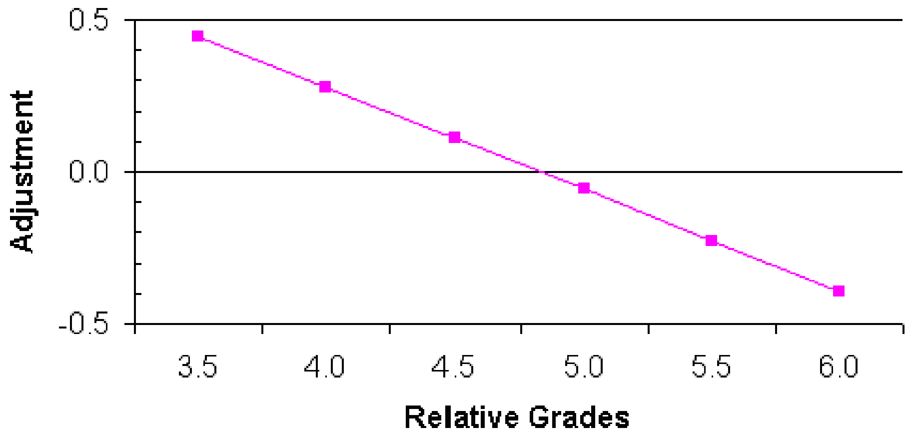

Figure 1: Example Adjustment for Relative Expected Grade.

I nstructional
A ssessment
S ystem

USE NO. 2 PENCIL ONLY

Fill in bubbles darkly and completely.
Erase errors cleanly.

FORM
A

Instructor ____________ Course ____________ Section ______ Date ____________

Completion of this questionnaire is voluntary. You are free to leave some or all questions unanswered.

	Excel-lent	Very Good	Good	Fair	Poor	Very Poor
1. The course as a whole was:	◯	◯	◯	◯	◯	◯
2. The course content was:	◯	◯	◯	◯	◯	◯
3. The instructor's contribution to the course was:	◯	◯	◯	◯	◯	◯
4. The instructor's effectiveness in teaching the subject matter was:	◯	◯	◯	◯	◯	◯
5. Course organization was:	◯	◯	◯	◯	◯	◯
6. Clarity of instructor's voice was:	◯	◯	◯	◯	◯	◯
7. Explanations by instructor were:	◯	◯	◯	◯	◯	◯
8. Instructor's ability to present alternative explanations when needed was:	◯	◯	◯	◯	◯	◯
9. Instructor's use of examples and illustrations was:	◯	◯	◯	◯	◯	◯
10. Quality of questions or problems raised by instructor was:	◯	◯	◯	◯	◯	◯
11. Student confidence in instructor's knowledge was:	◯	◯	◯	◯	◯	◯
12. Instructor's enthusiasm was:	◯	◯	◯	◯	◯	◯
13. Encouragement given students to express themselves was:	◯	◯	◯	◯	◯	◯
14. Answers to student questions were:	◯	◯	◯	◯	◯	◯
15. Availability of extra help when needed was:	◯	◯	◯	◯	◯	◯
16. Use of class time was:	◯	◯	◯	◯	◯	◯
17. Instructor's interest in whether students learned was:	◯	◯	◯	◯	◯	◯
18. Amount you learned in the course was:	◯	◯	◯	◯	◯	◯
19. Relevance and usefulness of course content were:	◯	◯	◯	◯	◯	◯
20. Evaluative and grading techniques (tests, papers, projects, etc.) were:	◯	◯	◯	◯	◯	◯
21. Reasonableness of assigned work was:	◯	◯	◯	◯	◯	◯
22. Clarity of student responsibilities and requirements was:	◯	◯	◯	◯	◯	◯

SAMPLE

Relative to other college courses you have taken:

	Much Higher			Average			Much Lower
23. Do you expect your grade in this course to be:	◯	◯	◯	◯	◯	◯	◯
24. The intellectual challenge presented was:	◯	◯	◯	◯	◯	◯	◯
25. The amount of effort you put into this course was:	◯	◯	◯	◯	◯	◯	◯
26. The amount of effort to succeed in this course was:	◯	◯	◯	◯	◯	◯	◯
27. Your involvement in this course (doing assignments, attending classes, etc.) was:	◯	◯	◯	◯	◯	◯	◯

28. On average, how many hours per week have you spent on this course, including attending classes, doing readings, reviewing notes, writing papers and any other course related work?
◯ Under 2 ◯ 6 - 7 ◯ 12 - 13 ◯ 18 - 19
◯ 2 - 3 ◯ 8 - 9 ◯ 14 - 15 ◯ 20 - 21
◯ 4 - 5 ◯ 10 - 11 ◯ 16 - 17 ◯ 22 or more

29. From the total average hours above, how many do you consider were valuable in advancing your education?
◯ Under 2 ◯ 6 - 7 ◯ 12 - 13 ◯ 18 - 19
◯ 2 - 3 ◯ 8 - 9 ◯ 14 - 15 ◯ 20 - 21
◯ 4 - 5 ◯ 10 - 11 ◯ 16 - 17 ◯ 22 or more

30. What grade do you expect in this course?
◯ A (3.9-4.0) ◯ B (2.9-3.1) ◯ C (1.9-2.1) ◯ D (0.9-1.1) ◯ Pass
◯ A- (3.5-3.8) ◯ B- (2.5-2.8) ◯ C- (1.5-1.8) ◯ D- (0.7-0.8) ◯ Credit
◯ B+ (3.2-3.4) ◯ C+ (2.2-2.4) ◯ D+ (1.2-1.4) ◯ E (0.0) ◯ No Credit

31. In regard to your academic program, is this course best described as:
◯ In your major? ◯ A distribution requirement? ◯ An elective?
◯ In your minor? ◯ A program requirement? ◯ Other?

Mark Reflex® forms by Pearson NCS MM89765-3 654321 ED06 Printed in U.S.A.
©1995, University of Washington - Office of Educational Assessment

Figure 2: The Office of Educational Assessment's Evaluation Form.

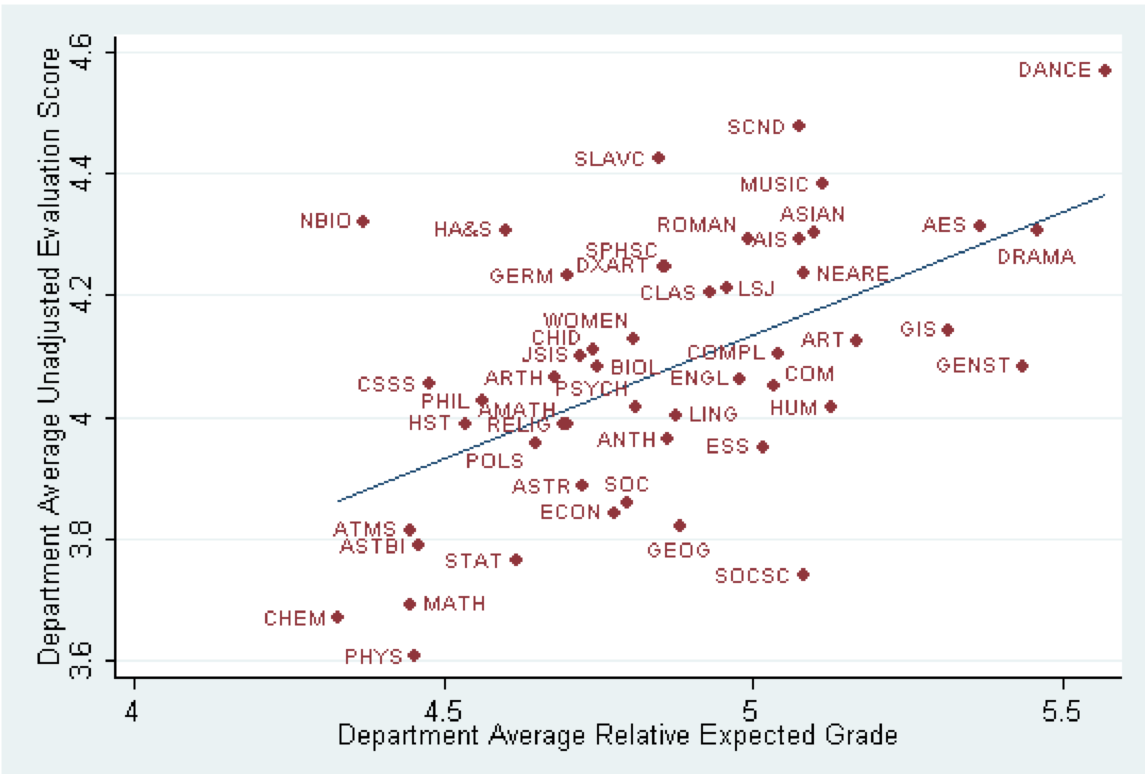

Figure 3: Average Relative Expected Grade and Average Evaluation Score by Department