Eastern Economic Journal, **2009**, **35**, (37–51)
© 2009 EEA 0094-5056/09
www.palgrave-journals.com/eej/

What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes

Michael A. McPherson, R. Todd Jewell and Myungsup Kim
Department of Economics, University of North Texas, PO Box 311457, Denton, TX 76203-1457, USA

Student evaluation scores are a standard component of the way colleges and universities assess the quality of an instructor's teaching for purposes of promotion and tenure, as well as merit raise allocations. This paper applies a feasible generalized least squares model to a panel of data from undergraduate economics classes. We find that instructors can "buy" better evaluation scores by inflating students' grade expectations. Class size and instructor experience are important determinants of evaluation scores in principles classes, but not in upper-level courses. Male instructors get better scores than females, and younger instructors are more popular than older ones. Certain other factors are also important determinants of evaluation scores. Our results suggest that an adjustment to the usual departmental rankings may be useful.

Eastern Economic Journal (2009) **35,** 37–51. doi:10.1057/palgrave.eej.9050042

Keywords: student evaluation; undergraduate economics

JEL: A0; A2

INTRODUCTION

Student evaluation of teaching (SET) at the college and university level and its determinants has been an area of active research for more than a half-century.[1] The large and growing literature in this area points to the importance of the role that SET scores have come to play in academic departments. For example, colleges and universities routinely use SET scores to assess the quality of an instructor's teaching for purposes of promotion and tenure. Furthermore, SET scores are often an important component in deliberations for merit or excellence raise allocations. While some strands of the literature in this area debate whether or not SETs should be of such central importance, the fact remains that these scores have been and continue to be used extensively. Understanding the determinants of SET scores may be of considerable interest and utility to instructors and to administrators.

Despite the breadth of the literature, much of the research has been unconvincing due to either data difficulties or statistical shortcomings. This paper takes advantage of an unusually large panel of data from 24 consecutive semesters comprising economics courses taught at a large public university. While McPherson [2006] analyzes a smaller portion of these data, his use of a fixed effects methodology precludes an examination of characteristics of instructors that are time-invariant. Instead, we use a random effects model estimated with feasible generalized least squares (FGLS). This enables an examination of instructor-specific, time-invariant characteristics such as gender and race. In addition, our method permits a proper accounting of unobservable effects specific to individual instructors. In the earlier literature there are only a small number of examples of efforts to tackle this important issue [Mason et al. 1995; Tronetti 2001; Isely and Singh 2005; McPherson 2006; McPherson and Jewell 2007].

A final area of interest involves the manner in which faculty members are ranked according to SET scores. Based on our estimation, we suggest at least two ways in which rankings could be usefully adjusted to account for extrinsic factors that might otherwise pollute the rankings. For example, if instructors can increase their evaluation scores by causing students to expect higher grades, departments may want to adjust rankings to eliminate the incentive to engage in such behavior. Similarly, if teaching intermediate-level theory classes means an instructor will receive lower evaluation scores than a colleague assigned to teach upper-level electives, then certain instructors may find themselves at a disadvantage when merit raise or tenure decisions are made. We show that such adjustments can lead to statistically significant changes in departmental rankings based on SET scores.

A REVIEW OF THE LITERATURE

A number of issues have been addressed in the literature. First and perhaps most obvious are concerns related to measurement. Do SETs actually measure effectiveness of instruction? Many have weighed in here: for example, Rodin and Rodin [1973] found a strong negative correlation between mean SET scores and performance on tests in calculus classes, indicating that less effective teachers get higher evaluations. Soper [1973] presents results showing that “students’ perceptions of their teachers’ abilities have no connection with what they learn” [p. 25]. However, White [1976], and Gramlich and Greenlee [1993], found a significant, if small, positive relationship between student performance and SET scores, perhaps indicating that more effective teachers do get better evaluations. Morgan and Vasché [1978] find that “…student evaluations [are useful] in indicating teaching productivity and identifying the most important teaching “attributes” for producing learning” [p. 126]. Marlin Jr. and Niss [1980] reach a similar conclusion. Although measurement issues are clearly important, this paper has a different focus. In any case, evaluating the determinants of SETs is a useful exercise, given the nearly universal use of them in US institutions of higher education.[2]

Another early focus of the literature was on whether or not instructors might “buy” higher SETs by entering into tacit agreements with students whereby higher student grades might be exchanged for higher SETs. For example, Voeks and French [1960], Kelley [1972], Costin et al. [1973] and others find either no evidence that instructors attempt to maximize their SETs in this manner, or very weak evidence. Villard [1973] offers intuitive reasons to expect that instructors will engage in such behavior, and McKenzie [1975] and Kau and Rubin [1976] provide a theoretical underpinning for such an expectation. Nichols and Soper [1972], Mirus [1973], and Dilts [1980] present empirical evidence that instructors may indeed attempt to buy higher SET scores. While these and many other works were central to the debate during the 1960s and 1970s, it was not until Seiver [1983] that any allowance was made for the likely endogeneity of expected grades. That is, while expected grade may affect how students evaluate teachers, it is also likely that quality of instruction (as may be measured by SETs) affects expected grade. As Seiver [1983] and others correctly note, the use of OLS methods to ascertain the determinants of SETs is quite likely to yield biased results. Seiver [1983] found that when a two-stage least squares (2SLS) procedure was employed to control for the endogeneity, there is no evidence that instructors are inflating grades (or grade expectations) in order to better their SET scores. Nelson and Lynch [1984] and Zangenehzadeh [1988] reached a similar conclusion using different data sets. However, Krautmann and Sander [1999] also considered the endogeneity question and reached the opposite conclusion. Still, endogeneity is evidently an important statistical matter in this area, and in that light it is surprising that some subsequent research ignored it altogether. For example, Aigner and Thum [1986] and McConnell and Sosin [1984] do not appear to recognize the issue at all. Stratton et al. [1994] and DeCanio [1986] make note of the issue in their papers, but fail to do anything about it. Clearly, failure to control for the endogenous nature of expected grades would render any findings suspect.

An important statistical issue involves the heterogeneity of instructors, and has barely received any consideration in the literature. While some heterogeneity, such as instructor gender, experience, education, rank, etc., can be controlled for, many others are unobservable. Yet failure to control for unobserved heterogeneity will lead to biased results. For the most part, previous research has struggled with this issue because of a paucity of time series observations on SETs. One obvious way to handle unobserved heterogeneity is to employ panel data methods. Only a small number of papers have explicitly sought to handle unobserved heterogeneity, including Mason et al. [1995], Tronetti [2001], Isely and Singh [2005], and McPherson [2006].

DATA

The data were obtained from the University of North Texas (UNT) Academic Records office and from the UNT Department of Economics. UNT is a comprehensive state university with more than 32,000 students. The Department of Economics has approximately 250 undergraduate majors but teaches many thousands of other students in its various course offerings.[3] The UNT Economics department is similar to programs at other large, state universities; thus, our data set is representative of that group, and our results should be broadly generalizable. These data represent 24 consecutive semesters between January 1994 and December 2005. Our data comprise 618 individual principles of economics classes taught by a total of 60 different instructors and 379 individual upper-level classes taught by 22 different instructors. Over this time period, there were a total of 76 instructors; we include data on 70 of them, 21 of whom are female and 24 of whom are non-white.[4] To test for the appropriateness of pooling the principles and upper-level courses, we conduct an *F*-test; the resultant *F*-statistic is 1.50 (43, 851), which is significant at the 2 percent level, indicating that pooling the two groups is inappropriate.[5] Thus, we analyze the two groups separately. The variables used in this study are discussed below, and summary statistics are given in Table 1.

Dependent variable

SET forms are distributed without announcement beforehand near the end of the semester and are anonymous. The Department of Economics uses an instrument that includes 25 questions, some of which are phrased in a positive and some in a negative manner. Given this instrument, there are many possible ways to measure quality of teaching. As the measure of SET, the Department of Economics computes

Table 1 Summary statistics

	Principles		*Upper-level*	
	N=619		*N=379*	
	Mean	*Standard deviation*	*Mean*	*Standard deviation*
evaluation	3.320	0.328	3.491	0.252
expgrade	2.906	0.263	3.222	0.289
pctmajor	0.009	0.018	0.409	0.225
pctfemale	0.532	0.099	0.497	0.146
response	0.678	0.124	0.694	0.135
macro	0.572	0.495		
elective3000			0.113	0.318
elective4000a			0.187	0.391
elective4000b			0.193	0.395
quantitative			0.082	0.274
size	69.05	36.86	26.92	10.44
twoday	0.417	0.493	0.441	0.497
threeday	0.507	0.500	0.174	0.380
male	0.544	0.498	0.755	0.431
white	0.708	0.455	0.794	0.405
experience	12.59	11.09	18.55	10.64
adjunct	0.850	0.358	0.172	0.377
age	35.96	7.277	41.30	9.341

an average of the student responses to the following four statements: “I would take another course that was taught in this way”; “The instructor did not synthesize, integrate, or summarize effectively”; “Some things were not explained very well”; and “I think that the course was taught quite well.” The average evaluation scores can range from one to four, with a four representing the best possible SET score. In this study, we utilize the Department’s chosen measure (*evaluation*).[6] The average SET score in principles classes is 3.32; the comparable statistic for upper-level classes is slightly higher at 3.49.

Independent variables

Following the literature, the determinants of SET scores are likely to fall into several categories. First are characteristics of the students in each class, which include the average expected grade in the course as reported by the student (*expgrade*), the proportion of students completing the evaluation questionnaire that major in economics (*pctmajor*), the proportion of students that is female (*pctfemale*), and the percentage of students enrolled in the class that participate in the evaluation exercise (*response*).[7] *Expgrade* is measured on a four-point scale, averaging 2.91 for principles courses in the data and 3.22 for upper-level courses. *A priori*, one would expect higher evaluation scores to be correlated with higher expected course grades.[8] The proportion of students majoring in economics may affect evaluation scores in that economics majors are presumably more interested in economics in general and, in addition, may be more likely to recognize quality teaching in economics. The gender composition of the respondents may impact SET scores if there are differences in the standards used by male and female students in evaluating teaching. The response rate may be an indicator of student enthusiasm for the course; in this case we might expect a higher response rate to cause higher evaluation scores.

Alternatively, *response* may be higher in courses in which attendance is perceived as particularly important. In this case the relationship between *evaluation* and *response* may be an inverse one. Finally, we include dummy variables representing the 24 semesters in order to control for changes in the composition and preferences of students over time.[9]

A second group of determinants of SET scores are characteristics of the course. We include a series of dummy variables indicating the type of the course. For principles classes, *macro* equals one if the course is a principles of macroeconomics section, with the excluded category being principles of microeconomics classes. Upper-level courses are divided into five categories: *elective3000* equals one if the course is a junior-level elective course; *elective4000a* equals one if the course is a senior-level elective course without an intermediate theory prerequisite; *elective-4000b* equals one if the course is a senior-level elective course that does have an intermediate theory prerequisite; and *quantitative* equals one if the course is a statistics or econometrics course. The excluded category for upper-level courses comprises intermediate-level theory courses that are required of all Economics majors, including Intermediate Microeconomics, Intermediate Macroeconomics, and Money and Financial Institutions.

The number of students in a given class also may be an important determinant of SET scores. Specifically, SET scores may deteriorate with increases in class size up to a point, after which students may perceive advantages to being relatively anonymous. As a result, we include both *size* and its square as regressors. The average class size for principles and upper-level courses is, respectively, approximately 69 and 27. Another important course-specific characteristic is the number of days per week that the course meets. This aspect is modeled with two dummy variables: *twoday* equals one if the course meets twice a week, while *threeday* equals one for classes that meet thrice weekly. As all courses in these data are three-hour courses, this is equivalent to controlling for the length of the class meeting on any given day. For example, 44.1 percent of upper-level courses meet twice a week, for one-and-a-half hours. The remainder meets three times per week for one hour (17.4 percent) or once per week for 3 hours (38.5 percent).

The third group of SET score determinants is instructor-specific characteristics. To control for unobservable characteristics, we take advantage of the longitudinal nature of the data and employ a panel data estimation approach. By “unobservable characteristics,” we mean those characteristics of the instructor that are either unobservable to the researcher or not quantifiable; these characteristics are assumed to be observable to students and, thus, have an impact on SET scores. For instance, the personality characteristics of an instructor may affect SET scores but cannot be included as regressors. Given the UNT data, we could use either a random-effects or a fixed-effects specification. Hausman tests for each sample indicate that the assumption of the random effects model concerning the orthogonality of the random effects and the regressors is appropriate for both the principles and upper-level samples. The chi-square statistic is 29.95 (33 degrees of freedom) for the principles sample and 35.49 (36 degrees of freedom) for the upper-level sample, both of which are insignificant at any conventional level. Thus, we cannot reject the null hypothesis of no correlation between the random effects and the regressors in either sample.[10] We choose a random effects model since it allows for the inclusion of time-invariant regressors.

To control for observable characteristics, we include the gender (*male*), race (*white*), total semesters of university teaching experience of each instructor (*experience*), whether the instructor is a teaching fellow or adjunct (*adjunct*), and the instructor's age (*age*).[11] Under the assumption that race and gender do not have an impact on teaching ability, an instructor's race and gender can still have an impact on SET scores if some bias exists in the evaluation process. For instance, research exists suggesting that students perceive female instructors differently than men.[12] SET scores are expected to increase with *experience*, since more time in the classroom should increase the quality of one's teaching. An instructor who is an adjunct or a teaching fellow (*adjunct*) has no research and limited service responsibilities.[13] Adjuncts and teaching fellows are hired for one purpose: teaching. In general, we expect that such faculty will have higher SET scores, all else equal.

Holding constant the effect of *experience*, we expect SET scores to deteriorate with *age*.[14] There are several reasons to expect such an effect. First, the effect may be due to reduced time spent on teaching activities relative to other tasks, such as research or administrative duties, as an instructor ages. Second, students may simply prefer courses taught by younger instructors. Third, as an instructor ages, she becomes further removed from her graduate education. Without additional training, an instructor's human capital, in terms of her knowledge of the current state of the discipline, will inevitably erode.

RESULTS

In the random effects model, individual-specific effects measuring unobservable instructor characteristics are modeled and estimated as being randomly distributed across instructors. Our random effects specification is given in equation (1).

$$evaluation_{ijt} = (\alpha + u_i) + X_{jt}\beta + Z_{it}\gamma + \varepsilon_{ijt} \tag{1}$$

The dependent variable is the SET score for each instructor i in course j at semester t. The instructor-specific constant, which is time-invariant, is the combination of a common constant term (α) and the instructor-specific effect (u_i). The matrix X_{jt} contains student-reported measures and course-specific variables for course j at semester t, the matrix Z_{it} contains instructor-specific information for instructor i at semester t, the vectors β and γ represent parameters to be estimated, and ε_{ijt} is a well-behaved, normally distributed error term.

Equation (1) is estimated separately for principles and upper-level courses using FGLS [Greene 2003, pp. 293–8], and the results are presented in Table 2.[15] We use a weighted estimation in each case since the error variances of SET scores will be larger for smaller courses; as weights, we use the number of students who filled out the instrument (*evalnumber*). The random effects model assumes that u is normally distributed with variance σ_u^2. FGLS allows the variance of u to vary across instructors, which is important since we have heteroskedasticity due to unbalanced panels. That is, principles instructors are observed on average for 10 courses, but some instructors have taught as few as two courses or as many as 80, while instructors of upper-level courses have taught on average 17 courses, with a range of between 2 and 41.

Principles courses

The results in Table 2 suggest that the student-reported measures in principles classes significantly affect SET scores. We find that the coefficient on *expgrade* is

Table 2 FGLS estimates

	Principles *N=618*	*Upper-level* *N=379*
constant	3.5172***	3.5582***
	(0.1474)	(0.1382)
expgrade	0.2714***	0.1036***
	(0.0314)	(0.0350)
pctmajor	1.0539**	−0.1976***
	(0.4325)	(0.0629)
pctfemale	0.2106***	−0.0413
	(0.0598)	(0.0579)
response	−0.0058	−0.1759***
	(0.0540)	(0.0637)
size	−0.00003	−0.0014
	(0.0001)	(0.0010)
twoday	0.0393	0.0303
	(0.0307)	(0.0194)
threeday	0.0070	0.0639***
	(0.0307)	(0.0255)
macro	0.0049	
	(0.0121)	
elective3000		0.1894***
		(0.0281)
elective4000a		0.2733***
		(0.0277)
elective4000b		0.1987***
		(0.0324)
quantitative		0.2880***
		(0.0449)
male	0.0940***	0.0659***
	(0.0197)	(0.0235)
white	−0.0065	0.1195***
	(0.0194)	(0.0239)
adjunct	0.0415**	0.1129***
	(0.0208)	(0.0285)
age	−0.0352***	−0.0128***
	(0.0021)	(0.0014)
experience	0.0202***	−0.0009
	(0.0013)	(0.0011)
log likelihood	201.057	176.535
chi^2 (degrees of freedom)	945.94***(36)	611.88***(39)

** Significant at 5 percent level.

*** Significant at 1 percent level.

Dependent variable = *evaluation* (weight = *evalnumber*) (Standard Errors in Parentheses).

positive and statistically significant. This means that principles instructors can "buy" higher scores by increasing the grade expectations of their students; specifically, inflating students' expected grade by one letter grade would cause an instructor's evaluation score to rise by 0.2714 points. As noted above, the past literature is quite unsettled on this point. Some researchers report similar findings to ours [Krautmann and Sander 1999]. Other researchers find no evidence of a causal link between expected grade and SET scores [Seiver 1983; Nelson and Lynch 1984]. Principles classes with a higher proportion of students declaring economics as their major assign instructors substantially higher evaluation scores. It also appears that classes comprising more female students give higher SET scores on average than classes with more male students.[16] The *response* rate, class *size*, and class meetings per week are not statistically significant determinants of *evaluation* for principles classes. Furthermore, there is no evidence that *evaluation* differs significantly between principles of macroeconomics and principles of microeconomics.

Turning to the instructor characteristics, Table 2 suggests that SET scores are strongly affected by observable characteristics. Instructor teaching *experience* is of considerable importance in determining *evaluation*. In particular, an additional semester of teaching experience is predicted to increase *evaluation* by 0.0202 points. The gender of the instructor also plays an important role in the determination of evaluation scores. *Ceteris paribus*, *male* instructors receive SET scores that are 0.094 points higher than their female colleagues. This finding suggests that students may perceive instructor quality differently according to instructor gender. However, we find that *evaluation* does not vary significantly by the instructor's race. As expected, we find that *adjunct* instructors have higher SET scores than tenure-track faculty. Finally, Table 2 demonstrates that *evaluation* deteriorates with *age*. Holding *experience* constant, it appears that UNT economics principles students give better SET scores to younger instructors. This may simply be a matter of the preferences of students; alternatively it may imply that older instructors are less interested in quality teaching or have less human capital.

Upper-level courses

Table 2 also presents FGLS results from the sample of upper-level classes. Inflating students' expected grade by one letter grade would cause an instructor's evaluation score to rise by 0.1036 points, an effect of lesser magnitude than that found in principles classes. It may be that students in upper-level courses are less likely to be "fooled" by their professors into thinking they will receive higher grades. Other student-reported measures also have significant effects on *evaluation*. Specifically, a 1 percent increase in the percentage of a class that is majoring in economics will lower an instructor's evaluation score by 0.1976 points. Note the difference in sign when compared to the principles sample; in the upper-level sample, the sign on *pctmajor* is negative, indicating that economics majors tend to be more critical of economics instructors than other students in upper-level courses, while they seem to less critical at the principles level. One could speculate that the quality of teaching in the principles courses lead majors to have certain expectations about economics instructors, and these students find that their upper-level instructors do not live up to their expectations. The proportion of the class that participates in the evaluation exercise (*response*) is inversely related to *evaluation*. In contrast to principles classes, there is no particular advantage to teaching a course with a high percentage of female students.

Similar to the result found in the principles sample, *size* does not play a significant causal role in the determination of SET scores. Instructors of upper-level classes that meet three times a week receive *evaluation* scores that are 0.0639 points higher than their colleagues who teach once a week. The type of course also plays a sizeable role in determining SET scores in upper-level courses. Instructors of the required theory courses receive substantially lower scores than instructors teaching any other sort of upper-level course. For example, instructors who teach quantitative courses can expect an evaluation score that is 0.288 points higher than instructors of theory courses. While not as quite as large, similar effects can be found for senior-level and junior-level electives.

A number of characteristics of the instructor also play significant roles in the determination of the evaluation score. Interestingly, instructor *experience* appears to have no significant effect on SET score for upper-level courses. This contrasts with principle courses, in which additional teaching experience improved an instructor's evaluation score. As was the case with principles classes, the instructor's gender appears to have an influence on *evaluation*. For upper-level classes, *male* instructors' SET scores are 0.0659 points higher than those of females, a slightly smaller effect that found in the principles sample. In contrast to the result found for principles classes, the instructor's race helps determine his or her SET score. Specifically, *white* instructors receive evaluation scores that are 0.1195 points higher than non-white instructors. This difference in racial effects across samples may be a reflection of the greater diversity among student populations taking principles classes relative to upper-level economics classes. *Adjunct* faculty members receive higher *evaluation* scores than their tenure-track counterparts; this effect is also observed in the principles data, but it is nearly three times as large in the upper-level sample. Finally, additional years of instructor *age* lead to a worsening of evaluation scores, similar to the situation with principles classes.

Adjusted rankings

Several researchers [Danielsen and White 1976; Mason et al. 1995; McPherson 2006] have suggested adjusting raw SET scores to eliminate the influence of factors that either could be manipulated by instructors to their advantage (e.g., expected grade) or that might be beyond an instructor's control (e.g., instructor's race or gender). For instance, a department could rank instructors based on predicted instructor-specific random effects, which would hold constant all observable effects. The random effects might be thought of as overall or longer-term indicators of instructors' teaching taking out the effects of expected grade, instructor gender, and all other observable effects. While it is tempting to think of these random effects as a measure of an instructor's underlying quality, there are some factors that may be part of the effect that have little to do with quality. For example, Hamermesh and Parker [2005] present evidence that students' perceptions of an instructor's physical beauty affect their evaluation scores. If it were the case that a large part of an instructor's random effect is the result of factors such as beauty, then an adjusted ranking based on random effects might not "improve" the rankings in any meaningful way.[17]

Another way to adjust rankings is to produce a predicted SET score for each instructor in each semester. That is, given the values of the explanatory variables for a given instructor in a given semester, we can produce a fitted value for the SET score of each instructor using the estimated coefficients presented in Table 2. This predicted value would not be influenced by the instructor-specific random effects; in addition, any of the explanatory variables can be removed from the adjustment in order to remove its influence. For example, a fitted value can be computed that assumes all students have the same grade expectations. A ranking based on this measure would effectively remove the advantages that might accrue to instructors who inflate grade expectations. Similarly, we could produce an adjusted rankings based on SET scores that are purged of any effect associated with gender or race.

Table 3 Adjusted rankings: overall average

Instructor	*Instructor gender*	*Instructor race*	*n*	*Raw ranking*	*Adjusted ranking: expgrade = sample mean*	*Adjusted ranking: male = 1 and white = 1*
A	M	W	42	1	1	2
B	M	W	7	2	2	3
C	F	W	12	3	3	5
D	F	W	12	4	6	9
E	M	W	19	5	9	10
F	M	W	33	6	5	4
G	F	W	23	7	4	6
H	M	W	45	8	7	7
I	M	W	39	9	8	8
J	M	NW	4	10	16	15
K	M	NW	38	11	12	1
L	M	W	10	12	13	16
M	M	NW	10	13	15	11
N	M	W	44	14	11	14
O	M	W	35	15	17	17
P	F	W	5	16	10	12
Q	M	NW	24	17	14	13

In Table 3, we present a ranking based on these types of adjustments averaging raw and predicted SET scores over all semesters, using the regressors in Table 2.[18] In the sixth column, we present a ranking after the effects of expected grade have been neutralized by assigning each instructor the mean *expgrade* of the relevant sample. While broadly similar to the ranking that is actually used (presented in the fifth column), for some instructors the change in ranking is rather dramatic. For example, Instructor J's ranking would dip from 10th to 16th, perhaps indicating that this instructor is inflating student grade expectations. On the other hand, Instructor P is ranked 16th in the raw rankings, but improves to 10th when the rankings are adjusted for expected grade. It appears to be the case that Instructor P's students expect lower-than-average grades and issue lower SET scores as a result. In column seven, we present a ranking adjusted for instructor gender and race. If all instructors were of the same gender and race, Instructor K, a non-white male, would see his ranking jump from 11th to first. Conversely, Instructors D (white female) and E (white male) drop from fourth and fifth to ninth and tenth, respectively.

While Table 3 provides information on how the rankings would change with certain adjustments to SET scores, it gives no indication as to the statistical significance of any change in the rankings. For instance, adjusting the rankings for expected grade leads to Instructors J and P trading places, but we cannot say whether or not this new ranking is statistically different from the unadjusted ranking. One way to test for significant changes in rankings associated with adjustments is to create confidence intervals for the adjusted (predicted) SET scores for each instructor and evaluate any overlap. Table 4 presents the predicted SET scores after adjustment as well as 95 percent confidence intervals.[19]

Generally, the point estimates and confidence intervals presented in Table 4 indicate that the adjustments do significantly change some instructors' rankings *vis-à-vis* their colleagues. For example, consider the movement of Instructor J and P when adjusting for grade expectations; given the point estimates and confidence intervals reported for these instructors, the SET score of Instructor P is significantly

Table 4 Adjusted rankings: estimates and 95% confidence intervals

Adjusted ranking: expgrade = sample mean				*Adjusted ranking: male = 1 and white = 1*			
Instructor	*Lower bound*	*Point estimate*	*Upper bound*	*Instructor*	*Lower bound*	*Point estimate*	*Upper bound*
A	3.6341	3.6584	3.6826	K	3.6334	3.6704	3.7074
B	3.5905	3.649	3.7075	A	3.6394	3.6632	3.687
C	3.5856	3.6296	3.6736	B	3.5747	3.6341	3.6935
G	3.5574	3.6228	3.6882	F	3.5631	3.5896	3.6161
F	3.5699	3.5976	3.6253	C	3.5328	3.5857	3.6386
D	3.544	3.5946	3.6451	G	3.5248	3.5813	3.6378
H	3.546	3.5829	3.6199	H	3.5432	3.5804	3.6175
I	3.5028	3.5363	3.5698	I	3.5029	3.5365	3.57
E	3.4782	3.5116	3.545	D	3.4704	3.5278	3.5852
P	3.4123	3.4892	3.566	E	3.4814	3.5153	3.5493
N	3.4485	3.4847	3.521	M	3.4452	3.4936	3.5421
K	3.3599	3.4046	3.4493	P	3.4172	3.4844	3.5517
L	3.3383	3.3832	3.4282	Q	3.435	3.4793	3.5236
Q	3.2169	3.2707	3.3244	N	3.4227	3.4555	3.4883
M	3.1642	3.2233	3.2823	J	3.3358	3.4152	3.4946
J	3.1004	3.1877	3.2749	L	3.363	3.4071	3.4512
O	3.1181	3.1769	3.2358	O	3.1216	3.1803	3.239

larger than that of Instructor J, and Instructor P jumps over Instructor J in the rankings. Now consider the same grade expectation adjustment for Instructors E and F, who are ranked fifth and sixth in the raw ranking. After adjustment, Instructor F's ranking increases slightly to fifth, Instructor E's ranking drops to ninth, and the associated confidence intervals do not overlap. Thus, Instructor E's ranking decreases significantly when compared to that of Instructor F. Similarly, we find some significant changes in the ranking when adjusting for race and gender. The most obvious change is Instructor K who significantly jumps over Instructors D, E, F, H, I, and J.

Although not reported here, one could compute predicted SET scores adjusting for whatever variables are thought to pollute the rankings. For example, our results indicate that principles instructors can expect higher SET scores when they teach classes with larger proportions of female students. Also, upper-level instructors with the bad luck to be assigned to intermediate theory courses can expect significantly lower evaluation scores. While we are not making the statement that all departments must make these adjustments, our results do suggest that such adjustments (or lack thereof) may have far-reaching implications for individual faculty members. Therefore, it may be prudent for departments to examine and discuss the potential costs and benefits of adjusting SET scores.

Conclusions and policy implications

While the issue of whether the SET process actually measures quality of teaching output (and therefore whether or not SET scores should be used to evaluate instructors) may never be settled conclusively, the fact remains that SET scores are important components of both the promotion and tenure and merit raise allocation processes at many US universities. As such, a better understanding of the factors that drive evaluation scores is a worthwhile goal. This paper uses an FGLS approach to examine a panel of data comprising 618 individual principles classes and 379 upper-level classes over 24 consecutive semesters. This approach represents one of the few attempts to properly account for individual-specific unobservable effects.

Several principal findings emerge. We find that by inflating grade expectations, instructors can increase their student evaluation scores in both principles and upper-level classes. In addition, among principles courses instructors of courses with high proportions of students who are economics majors or women fare notably better than their colleagues teaching courses with different student compositions. In upper-level classes, the proportion of students majoring in economics also matters; however, the effect is a negative one. In principles classes, evaluation scores decrease with class size, although an improvement is seen in large classes. Several characteristics of the instructor also affect SET scores. Instructor experience is an important determinant of SET score, but only at the principles level. Male and white instructors receive higher SET scores than their female and non-white counterparts in upper-level classes; in principles courses, gender is a significant determinant of SET scores, while race is insignificant. Finally, evaluation scores are negatively related to instructor age in our sample.

Many departments rank instructors according to SET scores, and these rankings are commonly used to inform decisions about merit-based raises and are often an important component in the promotion and tenure process. Our results suggest that departments may usefully consider adjusting rankings to account for factors that can be manipulated by instructors to their advantage (especially expected grade) and factors that are beyond the control of the instructor (for example, race and gender). Finally, we show that such adjustment to rankings can lead to statistically significant changes in SET score rankings, a result that has clear implications for promotion, tenure, and salaries.

Notes

1. Perhaps the earliest research in this area was Heilman and Armentrout [1936].
2. Becker and Watts [1999] discuss the fact that most departments of economics use SET scores as part of their faculty evaluation process.
3. The Economics Department also has a terminal Master's program with approximately 40 students enrolled.
4. The department only conducts the evaluation exercise during the regular academic year. Our data therefore exclude summer and Maymester courses. Six instructors taught fewer than two classes or in fewer than two semesters and are excluded in order to allow for the estimation of both semester-specific and instructor-specific effects.
5. Since the fixed effects panel data estimator will always be consistent, we use it to test for the appropriateness of pooling. The test involves running a fixed effects model combining the principles and upper-level groups. We include interactions between all the variables included in Table 2 and whether the course is at the upper level; in addition, we interact the instructor-specific fixed effects and whether the course is at the upper level, allowing both the fixed effects and the coefficients to vary over principles and upper-level courses. This estimation is followed by an *F*-test of the joint significance of the interaction terms. Full results of this test (sometimes called a "Chow Test") are available from the authors.
6. The Department's averages are actually scaled so that one represents the best score. Here, we use an inverted scaling for ease of discussion. In addition to the estimates presented in this paper, we estimate other models using the responses to individual questions (available from the authors). No important differences in the estimates using the different measures of SET are found, so we report the results using the Department's chosen measure. The entire instrument is also available from the authors.

7. An alternative measure might be the percentage of students in a class who report that the class is required. However, in our data set most classes are required. This is especially true at the principles level — more than 92 percent of students report that their principles class is required. As a result, specifications including a *required* variable (available from the authors) yield statistically insignificant coefficients. In a similar vein, researchers may also wish to consider the "discouraged-business-major" hypothesis [e.g., Salemi and Eubanks 1996]. That is, in some institutions some students choose to major in economics having failed to meet the admissions criteria for the business school. While standards for economics majors are not lower at UNT than those for business majors, researchers at schools that do have these issues may find it more appropriate to not use an explanatory variable such as *pctmajor*.
8. Past research on undergraduate SET scores indicates that expected grade may be endogenous [e.g., Seiver 1983 and Nelson and Lynch 1984]. Employing a Hausman test (available from the authors), we do not find evidence of endogeneity in either the principles or the upper-level samples.
9. Note that the semester dummies may also pick up other time-varying information, such as the composition of instructors.
10. Complete results of these tests are available from the authors. The use of random effects creates the possibility of omitted variables bias. Specifically, if we do not control for all the relevant factors as regressors, there might be an omitted variable included in u_i. This causes inconsistency if the omitted variable is correlated with regressors. However, our data allow us to control for many individual-specific factors, and the choice of the random-effects estimation is backed by the Hausman test; thus, our estimation results should be reliable.
11. *Experience* includes teaching experience prior to employment at UNT. Given the time span over which we observe SETs, one may be concerned about the race and gender composition of our sample over time. Each semester, non-white instructors make up an average of 26 percent of total instructors, with a minimum of 13 percent and a maximum of 38 percent. Also, the UNT data show that there are an average of 33 percent female instructors per semester, with a minimum of 19 percent and a maximum of 50 percent. We also find that female instructors have slightly more teaching experience than their male counterparts (11.4 semesters to 10.8 semesters), while non-white instructors have less teaching experience than their white colleagues (7.9 to 12.5 semesters).
12. The older literature is summarized by Feldman [1993]. An example from more recent research can be found in Hamermesh and Parker [2005] who find that evaluation scores of instructors are influenced by students' perception of instructor beauty, and that the effect of beauty on evaluation scores differs by instructor gender.
13. Teaching fellows cannot teach upper division courses, while adjuncts hold at least a Master's degree and can by rule teach any upper-level course as well as principles.
14. One might expect *age* and *experience* to be highly correlated. However, the correlation coefficients between *age* and *experience* are not extremely large: 0.72 for the principles sample and 0.45 for the upper-level sample. Furthermore, analysis of variance inflation factors are not indicative of high correlation. *Age* and *experience* are also not highly correlated with the semester dummies.
15. The semester dummies are included in all estimations in this paper, but the coefficients are suppressed for brevity. In all cases, the semester-specific effects are jointly significant and are available from the authors.
16. We check for the potential of differential effects of the gender composition of the class and the instructor; however, this interaction is not statistically significant and is not included in these results. Likewise, we estimate models with interactions between *male*, *white*, and student characteristics, finding no significant interactive effects.
17. The FGLS random-effects estimator does not produce an estimate of the individual-specific effect. Following Greene [2003, p. 296], we could use the mean of the differences between *evaluation* and its predicted value as an estimate of the instructor-specific (time constant) random effect.
18. These rankings include data from both principles and upper-level courses for tenure track faculty only. Rankings including all instructors are available from the authors.
19. In matrix form, equation (1) can be written as $y = W\delta + v$, where y is SET score, W is a set of regressors, and v is the composite error term. Let $Var(v) = \Omega$. The FGLS estimator $\hat{\delta}$ is $(W'\hat{\Omega}^{-1}W)^{-1}W'\hat{\Omega}^{-1}y$, and its variance–covariance matrix is $(W'\hat{\Omega}^{-1}W)^{-1}$. The variance–covariance matrix of the predicted SET score, $Var(\hat{y})$, is $W(W'\hat{\Omega}^{-1}W)^{-1}W'$. Adjusted SET scores are generated by replacing certain regressors with predetermined values. Let W_a be the adjusted regressor matrix. The variance–covariance matrix of adjusted prediction becomes $Var(\hat{y}_a) = Var(W_a\hat{\delta}) = Var(W_a (W'\hat{\Omega}^{-1}W)^{-1}W'\hat{\Omega}_v{}^{-1}) = W_a(W'\hat{\Omega}^{-1}W)^{-1}W_a{}'$. We use Stata [StataCorp 2005] to generate $Var(\hat{y}_a)$ based on the estimated variance–covariance matrix of the FGLS estimator using the regressors reported in Table 2. Since the number of courses that an instructor teaches is greater than one, an

instructor's overall SET score is an average of the relevant course SET scores. This means that the variance of predicted SET scores will include not only the variance of individual predictions but also covariances between predictions. Confidence intervals are created using the standard error of the predicted (adjusted) SET scores. Full details of the computation of the confidence intervals are available from the authors.

References

Aigner, D.J., and F.D. Thum. 1986. On Student Evaluation of Teaching Ability. *The Journal of Economic Education*, 17(4): 243–265.

Becker, W.E., and M. Watts. 1999. How Departments of Economics Evaluate Teaching. *American Economic Association Paper and Proceedings*, 89(2): 344–349.

Costin, F., W.T. Greenough, and R.J. Menges. 1973. Student Ratings of College Teaching: Reliability, Validity, and Usefulness. *The Journal of Economic Education*, 5(1): 51–53.

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

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

Dilts, D.A. 1980. A Statistical Interpretation of Student Evaluation Feedback. *The Journal of Economic Education*, 11(2): 10–15.

Feldman, K.A. 1993. College Students' Views of Male and Female Teachers: Part II — Evidence from Students' Evaluations of their Classroom Teachers. *Research in Higher Education*, 34(2): 151–211.

Gramlich, E.M., and G.A. Greenlee. 1993. Measuring Teaching Performance. *The Journal of Economic Education*, 24(1): 3–13.

Greene, W.H. 2003. *Econometric Analysis*, 5th ed. Upper Saddle River, NJ: Prentice-Hall Inc.

Hamermesh, D.S., and A. Parker. 2005. Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity. *Economics of Education Review*, 24: 369–376.

Heilman, J.D., and W.D. Armentrout. 1936. Are Student-Ratings of Teachers Affected by Grades? *Journal of Educational Psychology*, 27(March): 197–216.

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

Kau, J.B., and P.H. Rubin. 1976. Measurement Techniques, Grades and Ratings of Instructors. *The Journal of Economic Education*, 8(1): 59–62.

Kelley, A.C. 1972. Uses and Abuses of Course Evaluations as Measures of Educational Output. *The Journal of Economic Education*, 3(Fall): 13–18.

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

Marlin, Jr., J.W., and J.F. Niss. 1980. End-of-Course Evaluations as Indicators of Student Learning and Instructor Effectiveness. *The Journal of Economic Education*, 11(2): 16–27.

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

McConnell, C.R, and K. Sosin. 1984. Some Determinants of Student Attitudes Toward Large Classes. *The Journal of Economic Education*, 15(3): 181–190.

McKenzie, R.B. 1975. The Economic Effects of Grade Inflation on Instructor Evaluations: A Theoretical Approach. *The Journal of Economic Education*, 6(2): 99–106.

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

McPherson, M.A., and R.Todd Jewell. 2007. Leveling the Playing Field: Should Student Evaluation Scores Be Adjusted?. *Social Science Quarterly*, 88(3): 868–881.

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

Morgan, W.D., and J.D. Vasché. 1978. An Educational Production Function Approach to Teaching Effectiveness and Evaluation. *The Journal of Economic Education*, 9(2): 123–126.

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

Nichols, A., and J.C. Soper. 1972. Economic Man in the Classroom. *The Journal of Political Economy*, 80(5): 1069–1073.

Rodin, M., and B. Rodin. 1973. Student Evaluation of Teachers. *The Journal of Economic Education*, 5(1): 5–9.

Salemi, M.K., and C. Eubanks. 1996. Accounting for the Rise and Fall in the Number of Economics Majors with the Discouraged-Business-Major Hypothesis. *The Journal of Economic Education*, 27(4): 350–361.

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

Soper, J.C. 1973. Soft Research on a Hard Subject: Student Evaluations Reconsidered. *The Journal of Economic Education*, 5(1): 22–26.

StataCorp. 2005. *Stata Statistical Software: Release 9*. College Station, TX: StataCorp LP.

Stratton, R.W., S.C. Myers, and R.H. King. 1994. Faculty Behavior, Grades, and Student Evaluations. *The Journal of Economic Education*, 25(1): 5–15.

Tronetti, R.J. 2001. Does Class Size Matter? Evidence from Panel Data Estimation, Master's Thesis, University of Central Florida.

Villard, H.H. 1973. Some Reflections on Student Evaluation of Teaching. *The Journal of Economic Education*, 5(1): 47–50.

Voeks, V.W., and G.M. French. 1960. Are Student-Ratings of Teachers Affected by Grades? The Report of Three Studies at the University of Washington. *The Journal of Higher Education*, 31(6): 330–334.

White, R.A. 1976. Some Added Support Justifying Administrative Use of Student Evaluations of Teachers. *The Journal of Economic Education*, 7(2): 120–124.

Zangenehzadeh, H. 1988. Grade Inflation: A Way Out. *The Journal of Economic Education*, 19(3): 217–226.