Research in Economic Education

In this section, the *Journal of Economic Education* publishes original theoretical and empirical studies of economic education dealing with the analysis and evaluation of teaching methods, learning, attitudes and interests, materials, or processes.

PETER KENNEDY, Section Editor

Determinants of How Students Evaluate Teachers

Michael A. McPherson

Abstract: Convincingly establishing the determinants of student evaluation of teaching (SET) scores has been elusive, largely because of inadequate statistical methods and a paucity of data. The author uses a much larger time span than in any previous research—607 economics classes over 17 semesters. This permits a proper treatment of unobserved heterogeneity. Results indicate that instructors can buy higher SET scores by awarding higher grades. In principles classes, the level of experience of the instructor and the class size are found to be significant determinants of SET scores. In upper-division classes, the type of student and the response rate matter. In both types of classes, factors specific to courses, instructors, and time periods are important; adjustments of scores to remove these influences may be warranted.

Key words: class size, student evaluations of teaching, unobserved heterogeneity
JEL code: A22

An extensive literature surrounds the issue of student evaluation of teaching (SET) scores. Research in this area began as early as 1936 with Heilman and Armentrout's article in the *Journal of Educational Psychology* and has continued unabated. The quantity of research is indicative of the importance of SET in higher education. For better or for worse, it is now standard for universities to

Michael A. McPherson *is an associate professor of economics at the University of North Texas (e-mail: mcpherson@unt.edu).* *The author is grateful for the assistance of David Molina and Caesar Righton in assembling the data and the insightful suggestions of Jeffrey Rous and R. Todd Jewell. The suggestions of three anonymous referees and Peter Kennedy were especially valuable.*

expect departments of economics to evaluate faculty, at least in part according to their SET scores (Becker and Watts 1999). The findings of researchers in this area are varied and sometimes in opposition to each other.[1] Unfortunately, statistical shortcomings and problems related to the data themselves have hampered much of the work in this area. My results add to the literature by addressing a critical statistical problem that has plagued previous work: unobserved heterogeneity. Although a few efforts have been made to tackle this problem (e.g., Mason, Steagall, and Fabritius 1995; Tronetti 2001), each has had other statistical shortcomings, such as a failure to test for endogeneity or a lack of time-series data.

In this study, I tested for endogeneity and controlled for unobserved heterogeneity and used a much longer time period than any previous work—17 semesters from 1994 to 2002. In the model, I employed controls for instructor, course, and semester-specific effects. The results suggested some obvious ways that rankings of instructors by SET scores might be adjusted. Clearly this is an important issue given the importance that such rankings have in determining such things as promotion, tenure, and merit raises. I examined whether adjustments of this nature would significantly alter the rankings of instructors.

MODEL AND DATA DESCRIPTION

I obtained the data from the University of North Texas's (UNT's) Academic Records Office and from the Department of Economics.[2] The data covered the 17 semesters between August 1994 and December 2002 and comprised 607 individual undergraduate classes taught by a total of 35 different instructors. It is possible that the relationships between SET scores and the explanatory variables differ for introductory economics classes compared with upper-division classes. Earlier researchers with access to data from both principles and upper-division courses routinely pooled the classes together without conducting tests of the appropriateness of such a grouping (see for example, Danielson and White 1976; Aigner and Thum 1986; Isely and Singh 2005). I conducted an F test for the appropriateness of pooling together principles courses with upper-division courses; these tests indicated that it is inappropriate to pool the data,[3] and, as a result I present regression results separately for principles and upper-division observations.[4] The principles subsample comprised 360 classes taught by 28 individual instructors. There were 247 upper-division classes, with 20 individual instructors. Both samples excluded classes with fewer than 15 students and included only instructors teaching at least three classes over the 17-semester time frame.

The instructors distributed SET forms without announcement beforehand[5] near the end of the semester and the forms were anonymous. The form included 25 questions, some of which were phrased in a positive and some in a negative manner. The answers were on 4-point scale with a 1 indicating *strong agreement* and a 4 indicating *strong disagreement* with the question. Given this instrument, there are many possible ways to measure quality of teaching. In this research, I used as the dependent variable (hereafter referred to as EVAL) an average of four questions: "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."[6] A second dependent variable consisted of the average value from the last of these four questions. Because the results differed only slightly, the results from regressions involving the second dependent variable are not presented here, although any substantial differences will be noted. In principle, EVAL can range from 1 to 4, with a 1 representing the best possible SET score and a 4 representing, at least from the students' points of view, poor teaching. For the principles classes sample, the average score for the dependent variable was (1.86) in Table 1. It is not surprising that evaluation scores were better for instructors of upper-division classes (1.74).

Following the literature, the determinants of the SET score are likely to fall into several categories. First are characteristics of the students in each class; these include such measures as major (PCTMAJ), expected grade (EXPGRADE), and the proportion of students who completed the evaluation questionnaire (RESPONSE).[7] PCTMAJ measures the percentage of the class that is majoring in economics; the average was 39.2 percent for upper-division classes and under 1 percent for principles classes. Economics majors might be more favorably disposed toward economics classes and instructors and perhaps better able than nonmajors to evaluate their economics instructors' abilities to teach economics, so SET scores might be better in such classes. I collected data on expected grades (EXPGRADE) as part of the evaluation exercise, and measured the variable on the usual 4-point scale, averaging 2.88 (principles classes) and 3.22 (upper-level courses) for these data. This effect has been of particular interest in the literature. Isely and Singh (2005) argued that it is the difference between expected grade and the grades that students are accustomed to receiving that matters. That is, when the average grade expected by a given class is above the average grade point average (GPA) of the class before the semester began, higher SET scores may result. However appealing this variable may be intuitively, it is not clear that such a variable can be properly calculated. The average expected grade was calculated from the evaluation exercise and, as such, represented only those students who were present on the day of the exercise and who chose to participate. The average GPA of the students responding to the evaluation questionnaire was not known, however. Instead, it was the average GPA of all students registered for the course that was available. Given that the distribution of students who did not participate in the evaluation process was unlikely to be the same as that of students who did, the validity of such a variable was questionable at best. In any event, I calculated this surprise variable in a similar manner to Isely and Singh—as the difference between expected grade and overall GPA. As I note later, a series of Davidson-MacKinnon *J* tests (Davidson and MacKinnon 1981) indicated that, in general, it was more appropriate to use EXPGRADE than this surprise variable. I included the response rate, measured as the percentage of the students registered who actually participated in the evaluation process, as a way to control for possible selection bias. In addition, this variable may be indicative of student interest, in which case it would be reasonable to expect it to have a beneficial effect on SET scores. Alternatively, a high response rate might mean that a larger number of poorly performing students were evaluating their instructors. This might have detrimental effect on an instructor's SET score. In any case, RESPONSE averaged about 68

TABLE 1. Descriptive Statistics

Variable	Principles classes Mean (st. dev.)	Min.	Max.	Upper-division classes Mean (st. dev.)	Min.	Max.
EVAL	1.86 (0.40)	1.25	3.77	1.74 (0.42)	1.05	3.13
EXPGRADE	2.88 (0.24)	2.31	3.80	3.22 (0.29)	2.33	3.92
PCTMAJ	0.58 (1.12)	0.00	7.69	39.18 (22.57)	0.00	100.00
ONEDAY	0.08 (0.26)	0.00	1.00	0.38 (0.49)	0.00	1.00
TWODAY	0.42 (0.49)	0.00	1.00	0.39 (0.49)	0.00	1.00
THREEDAY	0.51 (0.50)	0.00	1.00	0.23 (0.42)	0.00	1.00
CLSIZE	82.34 (44.33)	19.00	318.00	32.96 (8.89)	16.00	53.00
EXPERIENCE: 1 TO 4 SEMESTERS	0.37 (0.48)	0.00	1.00	0.27 (0.44)	0.00	1.00
EXPERIENCE: 5–10 SEMESTERS	0.35 (0.48)	0.00	1.00	0.37 (0.48)	0.00	1.00
EXPERIENCE: 11+ SEMESTERS	0.27 (0.45)	0.00	1.00	0.36 (0.48)	0.00	1.00
RESPONSE (rate)	67.89 (12.51)	26.74	97.87	69.89 (13.76)	33.33	100.00
ECON1100: Micro	0.38 (0.49)	0.00	1.00			
ECON1110: Macro	0.62 (0.49)	0.00	1.00			
ECON3000: Contemp. issues				0.02 (0.13)	0.00	1.00
ECON3050: Consumer				0.04 (0.21)	0.00	1.00
ECON3150: Discrimination				0.06 (0.24)	0.00	1.00
ECON3550: Micro				0.15 (0.36)	0.00	1.00
ECON3560: Macro				0.12 (0.33)	0.00	1.00
ECON4020: Money and banking				0.15 (0.36)	0.00	1.00
ECON4100: Comp. systems				0.03 (0.18)	0.00	1.00
ECON4140: Managerial				0.03 (0.18)	0.00	1.00
ECON4150: Public finance				0.06 (0.25)	0.00	1.00
ECON4180: Health				0.03 (0.18)	0.00	1.00
ECON4290: Labor				0.03 (0.17)	0.00	1.00
ECON4440: Environmental				0.04 (0.19)	0.00	1.00
ECON4460: Ind. organization				0.01 (0.09)	0.00	1.00
ECON4500: Sports				0.01 (0.11)	0.00	1.00
ECON4510: History of thought				0.03 (0.18)	0.00	1.00
ECON4600: Development				0.03 (0.17)	0.00	1.00
ECON4630: Research methods				0.01 (0.09)	0.00	1.00
ECON4850: Trade				0.07 (0.25)	0.00	1.00
ECON4870: Econometrics				0.07 (0.25)	0.00	1.00
Sample size	360			247		

percent and ranged from about 27 percent to 100 percent. The response rate for principles classes was not significantly lower than that of the upper-division sample.

A second group of possible determinants of SET are characteristics of the course, such as the level of the course, the length of the class period, the number of students in the class, and so forth. In particular, I modeled the level of the course using a series of dummy variables. In the case of the principles data, the base category was principles of macroeconomics (ECON1110). As shown in

Table 1, 62 percent of the principles classes in the data were principles of macroeconomics, with the remaining 38 percent, principles of microeconomics. With respect to the upper-division courses, there were 20 different courses students could take. In the regressions involving upper-division classes, intermediate micro (ECON3550) served as the base category. The most common upper-division classes were intermediate macro (ECON3560), intermediate micro (ECON3550), and money and banking (ECON4020), the required courses for economics majors.[8]

Another characteristic of the course that I considered was the number of days per week that the course met. This aspect was modeled using two dummy variables, ONEDAY and THREEDAY. As all courses in these data were 3-credit hour courses, this was equivalent to controlling for the length of the class meeting on any given day. For example, 38 percent of upper-level courses met once a week; each meeting was 3 hours in duration. Principles classes were more commonly taught thrice weekly and, less commonly, meet once a week, compared to the upper-division sample. The base category in this case was classes that met twice weekly for $1^1/_2$ hours per lecture.

In many earlier contributions to the literature, researchers have studied the effects of class size on SET scores. Becker and Powers (2001) discussed the sample selection problem inherent in studies such as mine: Students who do not expect to be performing well in class and those who do not like their instructors are more likely to withdraw from a class than are other students. Although the data do not permit a comprehensive treatment of this issue, the findings of Becker and Powers suggest that the appropriate measure of class size is the enrollment at the beginning of the semester because class-size measures that are based on terminal enrollment or an average of initial and terminal enrollment are likely to be endogenous. In the present work, CLSIZE was defined as the number of students enrolled in the class at the beginning of the semester.[9] As class size increases, teaching methods must change. It may be reasonable to assume that students view larger class sizes in a negative manner.[10] The class size in the principles dataset ranged from 19 to 318, with an average of 82.3, whereas the average number of students in upper-level classes was just under 33, with a range from 16 to 53.

I used two dummy variables in an attempt to control for instructor experience. The first had a value of 1 if the instructor in question had between 5 and 10 semesters of experience, and the second took on a value of 1 if the instructor had 11 or more semesters of experience.[11] The base category, then, was courses taught by relatively inexperienced instructors.[12] Thirty-seven percent of principles classes were taught by relatively inexperienced instructors, and 27 percent had very experienced instructors. As one might expect, upper-division classes were more commonly taught by experienced instructors.

To control for the unobservable characteristics of the instructor, course, and semester, and to take advantage of the fact that $8^1/_2$ years of data were available, I used a panel approach, specifically a three-way fixed-effect model.[13] In addition to other explanatory variables, the fixed-effect model I included a dummy variable for each instructor, as well as a dummy variable for each semester. For the

present research, the equation of interest was as follows:

$$y_{itj} = \beta_1 + \alpha_i + \gamma_t + \lambda_j + \sum_{k=2}^{K} x_{kitj}\,\beta_k + \varepsilon_{itj}, \quad (1)$$

where α_i represents the fixed-effect specific to instructor i, γ_t represents the fixed-effect specific to semester t, λ_j represents the fixed-effect specific to course j, x_{kitj} includes the class, course, and instructor specific explanatory variables listed above, and ε_{itj} is assumed to be well-behaved.

Isely and Singh (2005) also used a fixed-effect model to examine the determinants of SET scores. However, because their focus was on differences in the way that a given instructor taught different courses, they used a one-way fixed-effect model to examine the variations of SET for a particular instructor, course, and section from the average SET for that instructor in that course as a function of similar deviations of expected grades and other control variables. This was equivalent to having a dummy variable for each course of each instructor. In this arrangement, the fixed-effect coefficient would amount to the intercept for a given instructor teaching a specific course. However, this specification sacrifices the ability to gauge the effect that teaching a particular course may have on SET scores regardless of who the instructor is (that is, factors intrinsic to the particular course but not specific to instructors). Furthermore, the Isely and Singh method did not allow an overall comparison of instructors. One could only say that one instructor rated better than another in a given course.[14]

There are reasons to believe that expected grade is an endogenous variable. Although an instructor's inflating of students' grade expectations might lead to the class assigning better average evaluation scores, if it is true that better teachers receive better evaluation scores, instructors with better evaluation scores will naturally have better performing students who expect higher grades. The empirical evidence on this sort of endogeneity is mixed: Seiver (1983) and Nelson and Lynch (1984) found endogeneity to be a problem, whereas Krautmann and Sander (1999) and Isely and Singh (2005) found the opposite. If EXPGRADE is endogenous, ordinary least squares (OLS) would yield biased and inconsistent parameter estimates. In such cases, these data should be analyzed using a two-stage least squares (2SLS) procedure. I carried out Hausman specification tests for each model to determine whether EXPGRADE was endogenous.

RESULTS

Principles Classes

Because tests for pooling data indicate that it is inappropriate to pool principles and upper-division courses, each subset of the data was considered separately. Hausman specification tests indicated that there was no evidence that EXPGRADE was endogenous, and, as a result, the use of OLS techniques was warranted. The regressions using data only from principles classes are presented in Table 2; there was a significant effect of expected grade on SET scores, with an increase in the average expected grade of 1 point on the usual 4-point scale causing an improvement

TABLE 2. Ordinary Least Squares Regression Results

Variable	Principles classes	Upper-division classes
CONSTANT	2.7962*** (13.572)	2.3189*** (7.276)
EXPGRADE	–0.3417*** (–5.748)	–0.2999*** (–3.556)
PCTMAJ		0.0035** (2.022)
ONEDAY	0.0013 (0.027)	0.0499 (0.805)
THREEDAY	–0.0221 (–0.726)	–0.0017 (–0.027)
CLSIZE	0.0008*** (2.703)	0.0012 (0.467)
EXPERIENCE: 5–10 SEMESTERS	–0.1002** (–2.117)	0.0301 (0.276)
EXPERIENCE: 11+ SEMESTERS	–0.1728* (–1.836)	0.1975 (1.011)
RESPONSE	0.0013 (1.262)	0.0029** (1.996)
Course Fixed Effects		
ECON1100: Principles of micro	–0.0359 (–1.148)	
ECON3000: Contemp. issues		–0.1633 (–0.754)
ECON3050: Consumer		–0.0802 (–0.354)
ECON3150: Discrimination		–0.1641 (–1.422)
ECON3560: Macro		0.0562 (0.412)
ECON4020: Money and banking		–0.0661 (–0.408)
ECON4100: Comparative systems		0.0885 (0.281)
ECON4140: Managerial		0.0614 (0.353)
ECON4150: Public finance		–0.1111 (–0.968)
ECON4180: Health		–0.0745 (–0.447)
ECON4290: Labor		–0.5279*** (–3.241)
ECON4440: Environmental		0.0583 (0.369)
ECON4460: Ind. organization		0.0026 (0.010)
ECON4500: Sports		–0.4618** (–2.116)
ECON4510: History of thought		–0.1251 (–0.830)
ECON4600: Development		–0.1149 (–0.683)
ECON4630: Research methods		–0.3671 (–1.393)
ECON4850: Trade		–0.2209* (–1.742)
ECON4870: Econometrics		–0.3513 (–1.077)
Sample size	360	247
$\bar{R}^2$	0.779	0.646
F statistic	25.360	8.230

Notes: t statistics are in parentheses. *significant at a two-tailed Type I error level of .10. **significant at a two-tailed Type I error level of .05. ***significant at a two-tailed Type I error level of .01. Dependent variable is measured on a scale of 1 to 4, with lower scores representing better teaching evaluation scores. Estimates of the instructor- and semester-specific fixed effects are available from the author upon request.

in SET scores of about 0.34 points. The implication was that at the introductory level better teaching evaluation scores can be bought by instructors causing students to expect higher grades. The magnitude of the coefficient was comparable to that reported by Isely and Singh (2005) in their study of classes at Grand Valley State University but smaller than Dilts (1980) found at Ball State University or Krautmann and Sander (1999) at DePaul University.

Several characteristics of particular classes were important determinants of evaluation scores.[15] The number of days per week the class met was not an

important determinant of SET scores in either a statistical or an economic sense. This finding was somewhat surprising given that several earlier studies (Nichols and Soper 1972; Nelson and Lynch 1984; Isely and Singh 2005) found such effects to be important. This difference may be indicative of heterogeneity across universities or that these earlier studies failed to account for the several sources of unobserved heterogeneity considered here. Class size had a significant effect on SET scores; a one-student increase in class size caused evaluation scores to rise (worsen) by 0.0008 points. To understand this finding, consider two classes that are identical except that the first is average in terms of class size (82 students), and the second is one standard deviation greater than the mean (127 students). The evaluation score for the former would be 0.036 points better than the latter. The magnitude of this effect is similar to that found by Danielsen and White (1976) using data from the University of Georgia but smaller than that found by Krautmann and Sander (1999) and by Isely and Singh (2005). This improvement would have some effect on the rankings of instructors and is evidence that smaller class sizes are to be preferred in this regard.

Experience was of considerable importance in determining SET scores. In particular, instructors with between 5 and 10 semesters of experience had SET scores that were about 0.10 points better than instructors with less than 5 years' experience, *ceteris paribus*. Similarly, instructors with 11 or more semesters of experience had SET scores that were 0.17 points better than their less-experienced colleagues. The response rate was not a significant determinant of SET scores.

Fixed effects were generally important, with instructor-specific fixed effects especially so.[16] In particular, there was no significant difference between principles of microeconomics and principles of macroeconomics sections. However, the majority of the instructor-specific effects were different from zero in a statistical sense and were large in magnitude. As discussed later, it may be useful to consider these coefficients to be longer term measures of teaching quality, and, as such, instructors could be ranked accordingly. The best score in this regard belonged to instructor 3, and the highest (worst) score was associated with instructor 18. An F test of the null hypothesis that the instructor-specific fixed effects were jointly zero can be rejected at the 99 percent confidence level.[17] Only three of the semester-specific fixed-effects coefficients were statistically significantly different from zero. Nevertheless, an F test of the null hypothesis that the semester-specific fixed-effects were jointly zero can be rejected at a 99 percent confidence level.[18]

The regression that used the alternative dependent variable noted above produced very similar results, with only one notable difference. The estimated coefficient on the dummy representing 11 or more years of experience, although similar in magnitude to the regression reported in Table 2, was not statistically significant.

Upper-Division Classes

As was the case with the principles regressions, Hausman specification tests for the upper-division courses indicated that EXPGRADE was not endogenous, and so, once again, a simple one-stage FEM would be the appropriate specification.

First, the data allowed us to reject the hypothesis that the coefficient on EXPGRADE was zero (Table 2). The magnitude of the coefficient was comparable to that in the principles regression. This result was in opposition to Seiver (1983) and Nelson and Lynch (1984) but in accord with Isely and Singh (2005). Evidently, there was a significant negative relationship between EXPGRADE and SET score, implying that instructors might be able to increase their evaluation scores by inflating grade expectations.

Classes containing high proportions of economics majors seemed to be more critical of instructors than other classes. An increase in the percentage of the class that was majoring in economics worsened SETs scores by about 0.004 points. Instructors of a class made up exclusively of economics majors would receive evaluation scores that were about 0.20 points worse than a teacher of an identical class in which only half were majors. This result was somewhat surprising, because one might have reasonably expected nonmajors to be less appreciative of instruction in economics classes. It may be the case that nonmajors were more likely to have elected to take the class out of interest, whereas economics majors were more likely to have been required to take a given economics class. As was the case with principles classes, the number of days per week that an upper-division class met had no evident effect on SET scores once other factors were considered.

It is interesting to note that upper-division class size had no apparent effect on evaluation scores, unlike the situation with principles classes. This might be because there was much less variation in class sizes in upper-division courses, which ranged in size from 16 to 53 students, whereas principles classes ranged from as small as 19 to as large as 318. Unlike the principles case, instructor experience in upper-division teaching did not affect SET scores. Because the vast majority of upper-division classes were taught by faculty members holding doctoral degrees, students in a given class might perceive their instructor to be an expert regardless of the level of the instructor's experience.

The response rate was a statistically significant determinant of SET scores. Classes in which a higher proportion of students participated in the evaluation process tended to be more critical of their instructors, with each additional percentage point of attendance associated with a worsening of SET scores of around 0.003 points. Although this effect had statistical significance, it was rather small in magnitude.

Table 2 also reveals the extent to which the course fixed-effects were significant determinants of SET score. SET scores were significantly different from the base category (intermediate microeconomics) for only a few particular upper-division classes, but the coefficients for these were rather large. For example, teachers of the labor economics class (ECON4290) had SET scores approximately 0.53 points better than instructors of intermediate microeconomics. An effect of a similar magnitude existed for the sports economics (ECON4500) class and, to a lesser extent, international trade (ECON4850) and research methods (ECON4630).[19] Twenty different instructors taught upper-division courses during the 17 semesters spanned by the data. As was the case with the principles regressions, the majority of the instructor-specific dummies were significantly different

from zero and large in magnitude, indicating that factors peculiar to instructors form an important part of SET scores. These coefficients ranged in magnitude from about 0.46 to –0.37.[20] Finally, several semester-specific fixed effects were significantly different from zero.[21] Together, these findings suggested that much of what students considered when evaluating teachers involved difficult-to-measure aspects of the instructor, and to a lesser extent, the semester in which the course was taught and the course itself.

Once again, the specification that employs the alternative dependent variable produced very similar results to those reported in Table 2. The only important differences involved the statistical significance of the course-specific fixed-effects. In the alternative specification, teaching economics of discrimination (ECON3150) significantly improved an instructor's SET score, whereas teaching international trade (ECON4850) did not have a statistically significant effect.

Grade Surprise: Evaluating the Isely and Singh Variable

Before leaving this topic, it is useful to compare further my results with those of Isely and Singh (2005). Despite the concerns already noted about the variable, I constructed a grade surprise variable similar to that suggested by Isely and Singh. Following Isely and Singh, I carried out a series of Davidson-MacKinnon (1981) *J* tests to determine whether the specifications using EXPGRADE (model I) were superior to those employing the grade surprise variable (model II). The first step of the test involved estimating model II and calculating the fitted values from that regression. Model I was then estimated, with the fitted values from the first regression included among the regressors in the second regression. If the coefficient on this fitted value was found to be significantly different from zero, the implication was that model I was not the correct specification (otherwise it was). The second step was to reverse this procedure, estimating model I in the first step. Should the estimated coefficient on the fitted value be significantly different from zero, model II was the preferred specification. Logically, this means that the Davidson-MacKinnon *J* test might indicate that model I was clearly superior to model II, that model I was clearly inferior to model II, or that it was inconclusive. This latter finding would emerge if the coefficients on the fitted values from both parts of the test were found to be either significantly different from zero or both not significantly different from zero.

For the principles regressions, the Davidson-MacKinnon *J* test was inconclusive. The *t* statistic from the first step was 2.216; this implied that model I (the model with EXPGRADE) was not the correct specification. However, the *t* statistic from step two was 6.190, implying that model II (the model with the grade surprise variable) was also not the preferred specification.[22]

The Davidson-MacKinnon *J* test involving the upper-division classes indicate that the model that used EXPGRADE was preferable to that using the Isely and Singh variable. The *t* statistics for the first step was –0.859, implying that model I was the "true" model. The *t* statistic for the second step was 3.488, implying that the grade surprise specification was not preferred.

In short, for upper-division classes, Davidson-MacKinnon *J* tests indicated that specifications using EXPGRADE were preferred to those using the Isely and Singh variable. For principles classes, the surprise variable was not obviously preferable. For this reason, as well as because of the serious concerns noted earlier regarding the manner in which a grade surprise variable was calculated, EXPGRADE was used in all specifications in this article.[23]

ADJUSTING RANKINGS OF INSTRUCTORS

Several researchers in this area (Danielsen and White 1976; Mason, Steagall, and Fabritius 1995) 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 (such as type of course). The model presented above suggests at least two adjustments: a ranking based on the magnitude of the estimated fixed-effects coefficients, and a ranking based on an adjustment of each semester's raw SET score that accounts for extrinsic influences.

Fixed-effect Rankings

Instructors could be ranked according to their fixed-effect coefficients. In essence, an instructor's coefficient is the amount by which his or her intercept varies from the overall intercept that is common to all instructors. For example, for instructors of principles classes, the smallest and largest instructor-specific coefficients were –0.450 and 0.991, respectively (Table 3). Given the overall constant of 2.796 and a semester-specific fixed-effect of –0.207 in the fall 1994 semester, this implied that the fall 1994 intercept for the first instructor was 2.139, and for the second was 3.580. A comparison of these numbers held constant all observable effects as well as time-specific effects. It may be appropriate to think of these numbers as longer term measures of instructor quality. For example, the instructor 22's average SET score over all semesters in principles classes was 1.674. Out of the 28 principles instructors, this particular instructor would rank as seventh best. However, when ranked according to the fixed-effect coefficient, instructor 22 falls to the 12th position.

The rankings based on average SET score and the fixed-effect coefficients were relatively highly correlated, with Spearman's rank correlation coefficients equal to at least 0.95 for principles classes and at least 0.88 for upper-division classes (both were significantly different from zero in a statistical sense). This reflected the fact that the use of the fixed-effect coefficient changed an instructor's ranking by two or fewer positions in about half the cases. Still, certain instructors would be affected by the use of ranking based on the fixed-effect coefficients in a dramatic fashion. As previously noted, among principles instructors, instructor 22 was an example of a person who would see his or her ranking fall dramatically if fixed-effect coefficients were used to construct rankings. Instructor 44 was an example of a principles

TABLE 3. Comparison of Rankings Based on Average SET Score and Fixed-Effect Coefficients

	Principles instructors				Upper-division instructors			
Instructor	Average of EVAL (st. dev.)	Ranking based on average of EVAL	Fixed-effect coefficient (st. error)	Ranking based on fixed-effect coefficient	Average of EVAL (st. dev.)	Ranking based on average of EVAL	Fixed-effect coefficient (st. error)	Ranking based on fixed-effect coefficient
α_3	1.447 (0.111)	1	–0.450 (0.082)	1				
α_8	2.222 (0.269)	22	0.354 (0.031)	24				
α_7					1.989 (0.126)	17	0.462 (0.162)	19
α_9	1.518 (0.076)	2	–0.287 (0.040)	5	1.334 (0.158)	2	–0.133 (0.208)	8
α_{11}	1.831 (0.090)	18	0.003 (0.049)	18	1.901 (0.079)	15	–.016 (0.127)	12
α_{14}	1.772 (0.252)	15	–0.052 (0.051)	16	1.477 (0.323)	8	–0.120 (0.120)	9
α_{15}	1.547 (n.a.)	4	–0.356 (0.122)	2				
α_{16}	2.261 (0.047)	24	0.349 (0.099)	23				
α_{18}	2.706 (0.237)	27	0.991 (0.151)	28	2.224 (0.196)	19	0.321 (0.102)	17
α_{20}	2.048 (n.a.)	19	0.184 (0.121)	21				
α_{21}					1.646 (0.124)	12	0.029 (0.153)	13
α_{22}	1.674 (0.138)	7	–0.108 (0.115)	12	1.460 (0.247)	5	–0.290 (0.091)	3
α_{23}	2.826 (0.401)	28	0.855 (0.071)	27				
α_{24}	1.645 (0.162)	6	–0.309 (0.062)	4	1.543 (0.125)	9	–0.374 (0.273)	1
α_{25}	1.751 (0.171)	11	–0.136 (0.069)	9				
α_{26}	1.753 (0.084)	14	–0.127 (0.064)	11				
α_{27}					1.889 (0.343)	14	0.141 (0.078)	15

α_{28}	2.600 (0.409)	26	0.648 (0.095)	26				
α_{31}	1.689 (0.127)	8	–0.127 (0.097)	10				
α_{32}	2.111 (0.203)	21	0.148 (0.061)	20				
α_{33}	1.716 (n.a.)	10	–0.189 (0.106)	7				
α_{36}	1.813 (n.a.)	16	–0.076 (0.144)	14	1.441 (0.350)	4	–0.215 (0.261)	4
α_{37}	1.751 (0.208)	12	–0.027 (0.080)	17	2.057 (0.340)	18	0.091 (0.122)	14
α_{38}					1.966 (0.339)	16	0.525 (0.202)	20
α_{41}	2.061 (0.263)	20	0.105 (0.107)	19				
α_{42}					1.414 (0.174)	3	–0.056 (0.133)	10
α_{44}	1.752 (0.083)	13	–0.189 (0.090)	8				
α_{47}	1.713 (0.148)	9	–0.094 (0.036)	13	1.595 (0.087)	11	–0.054 (0.123)	11
α_{49}	2.333 (0.285)	25	0.435 (0.064)	25				
α_{50}	2.243 (0.289)	23	0.330 (0.116)	22	2.239 (0.313)	20	0.395 (0.120)	18
α_{52}					1.472 (0.202)	7	–0.206 (0.174)	5
α_{53}	1.534 (0.146)	3	–0.277 (0.045)	6	1.309 (0.106)	1	–0.342 (0.169)	2
α_{54}	1.586 (0.080)	5	–0.332 (0.090)	3	1.470 (0.076)	6	–0.156 (0.131)	7
α_{55}	1.829 (0.182)	17	–0.074 (0.061)	15	1.586 (0.340)	10	–0.185 (0.226)	6
α_{61}					1.651 (0.349)	13	0.213 (0.217)	16
Pearson's ρ			0.953				0.881	
t statistic			16.127				8.552	

instructor who would presumably prefer the fixed-effect rankings. For upper-division classes, instructor 24 jumped from 9th place (out of 20) in the average SET score, ranking to the best score, if fixed-effect coefficients rankings were used.

In short, these rankings were similar, but there were a few substantial differences. Ranking instructors by their fixed-effect coefficients may give a more complete picture of relative teaching quality in the sense that it takes into account factors that may be beyond an instructor's control.

Semester-by-Semester Rankings

If the principal interest were in adjusting scores for each semester, a comparison of the fixed-effect coefficients would not serve. There is one fixed-effect coefficient for each instructor based on information from all semesters. Most universities evaluate faculty each semester, and a given semester's performance might be better or worse than the overall trend for that instructor. As an alternative, suppose that for a particular semester, I start with the raw SET score but remove the influence of the observable extrinsic influences. For example, the results above suggest that the instructor can leverage better SET scores by causing students to expect higher grades. An alternative ranking could remove the rewards for such behavior. Other variables cause an instructor's evaluation score to worsen (for example, teaching an upper-division class with a large percentage of economics majors); it would be useful to compensate for such penalties. Mathematically, this adjustment could be represented as follows:

$$\tilde{y}_{itj} = y_{itj} - \hat{\lambda}_j - \sum_{k=2}^{K} x_{kitj}\,\hat{\beta}_k, \tag{2}$$

where $\tilde{y}_{itj}$ is the adjusted SET score, y_{itj} is the official SET score, $\hat{\lambda}_j$ is the estimated fixed-effect for course j, and $\hat{\beta}_k$ represents estimated parameters from equation (1). More experienced instructors should be allowed to receive whatever benefit accrues to them in the form of better evaluation scores. For this reason, the experience dummies were not used in the adjustment in the regressions. The adjustment did take into account the effects of expected grade, number of days per week the class meets, class size, response rate, and (in the case of upper-division classes) percentage of the class that was majoring in economics. In essence, equation (2) produces a ranking stripped of factors that can be manipulated by instructors to their advantage[24] and other factors beyond instructors' control but allows experience and otherwise unobservable instructor-specific effects to remain.[25]

Rankings based on the official or raw SET scores and the rankings based on the adjustment were generally relatively highly correlated, with Pearson's rank-correlation coefficients for principles classes ranging from 0.736 to 0.982 and for upper-division classes, from 0.709 to 0.964. With few exceptions, the official ranking and the adjusted ranking for any particular semester had a Pearson's correlation coefficient of at least 0.8.

Despite the high degree of correlation, important differences were caused by the adjustment of the rankings. First, in most semesters, there were faculty members who moved up or down several positions in the rankings after adjustment,

although in many cases, the adjustment caused movement of only one position in the rankings, if any. Instructor 24's rankings in principles classes in the spring of 1998 illustrates this point. This individual was rated as the seventh best teacher of principles classes out of 11 instructors if the rankings based on raw SET scores were employed. His or her position rose to the top spot if the rankings were adjusted. When considering instructors of upper-division classes, the adjustment once again worked to the benefit of instructor 24 in the spring 1995 semester—he or she moved from fifth to first place out of 11. Other instructors were affected (for better or worse) by this adjustment. Even a movement of one position could have important implications for a particular faculty member in personnel decisions, the allocation of merit-raise money, and the general esteem of colleagues.

A second and related point involves whether, over the course of many semesters, a particular faculty member is consistently over- or under-valued by the official rankings. If an instructor's official SET score ranking is either consistently above or consistently below his or her corrected ranking, over time some inequities might develop. To explore this possibility, I averaged the rank based on the raw SET score of a given instructor over all semesters in which he or she had taught and compared that with that instructor's average rank based on the corrected scores. In most cases, the correction did not have a large effect on an instructor's average ranking. That is, in most cases, an instructor might see his or her ranking change in the instructor's favor in one semester, but to his or her detriment in other semesters, largely balancing out over time. However, this was not the case for every instructor. For example, of the 12 semesters that instructor 24 has taught principles classes, in 8, the adjusted rankings were in the instructor's favor, and in 3, the adjusted ranking represented no change from the ranking based on the official SET score. In only one semester would this instructor's position in the rankings be adversely affected by the proposed adjustment. If rankings were adjusted each semester, instructor 24's average rank would be nearly two positions higher than is the case at present. The differences are even more important for faculty members teaching upper-division classes. Instructor 14, for example, would see his or her average ranking fall from 4.4 under the official ranking to 6th best under the adjustment. Presumably, over the years covered by these data, anyone evaluating faculty teaching based on an instructor's SET scores would have fairly consistently undervalued instructor 24's contribution and would have overvalued that of instructor 14. Quite a number of other instructors would be affected in a similar manner.

CONCLUSIONS

Even in the unlikely event that SET scores contain no information about the quality and effectiveness of teaching, the fact remains that they are widely used by instructors and administrators to evaluate teaching. This alone makes a better understanding of the determinants of SET scores worth pursuing.

Efforts to isolate the variables explaining SET scores have been made for more than 40 years. Unfortunately, early efforts suffered from one or more serious shortcomings in the statistical methods used, and all research has been hampered

to some extent by the unavailability of data from more than two or three consecutive semesters. This research is an effort to correct the previous problems and examine more completely the determinants of SET scores using a fixed-effect model. This specification deals appropriately with the unobserved course-specific, instructor-specific, and semester-specific effects that may affect SET scores. I also tested for endogeneity of expected grade. The data cover $8^1/_2$ academic years and so offer a unique glimpse into how the passage of time might affect SET scores.

Statistical tests indicate that it is inappropriate to pool principles and upper-division classes when examining SET scores, a finding that future research should take into account. A principal empirical finding involves the evidence regarding the possible contamination of SET scores by instructors attempting to buy better SET scores by raising grade expectations. In particular, higher expected grades do lead to significantly better SET scores among both principles and upper-division classes. This issue has been hotly debated in the literature, and this debate will surely continue. In any case, this finding underlines the importance of adjusting instructor rankings to remove any such effect.

The results of the present research also demonstrate that class size may affect SET scores, at least at the principles level. This finding indicates that teaching smaller classes results in better SET scores, *ceteris paribus*. If it is true that better SET scores are correlated with measures of student learning, this result reinforces the commonly held view that teaching is most effective in relatively small class sizes. In addition, certain other student and course attributes, such as the percentage of the class that is majoring in economics and the response rate, significantly influence SET scores in upper-division classes.

Unobserved course-specific effects are important determinants of SET scores in upper-division classes, with instructors of labor economics, sports economics, and international trade receiving better evaluation score than their colleagues, *ceteris paribus*. It is perhaps not surprising that there are no distinctions in principles classes between SET scores of instructors teaching microeconomics and those teaching macroeconomics.

Experience of instructors seems to have an important relationship with SET scores in principles classes, although it appears to be unimportant in upper-division classes. It is also important to note that the unobservable characteristics of instructors and to a lesser extent semesters (as captured by the fixed-effect coefficients) are typically large in magnitude and statistically significant. That is, these unobservable effects have at least as strong an influence on a typical instructor's SET scores as all other effects combined.

Adjusting rankings of instructors to account for influences beyond their control yields rank orderings that are relatively highly correlated with rankings based on raw or official SET scores. Nevertheless, important differences exist for certain faculty members. Given that many colleges and universities use rankings of instructors by SET scores as factors in promotion and tenure decisions, other personnel decisions, and the allocation of merit raises, the results presented here suggest that rankings adjustments may be appropriate and long overdue.

NOTES

1. An extensive review of this literature is available from the author on request.
2. Department of the Economics at UNT is part of the College of Arts and Sciences. However, many students from the College of Business take economics classes, and a one-degree program is jointly administered by the two colleges.
3. The value of the *F* statistic in this case is 2.00, and because the critical value of the test at the 95 percent confidence level is 1.75, I rejected the null hypothesis that pooling is appropriate.
4. It may not be appropriate to pool upper-division classes. Data constraints prevented this issue from being tested, but I presume that the inclusion of course-specific dummy variables dramatically lessens this potential problem.
5. Siegfried and Vahaly (1975) presented evidence that announcing evaluations in advance does not introduce any particular bias.
6. Because the second and third questions are phrased in a negative manner, these were rescaled by subtracting each from 5. Each question received equal weight.
7. The gender composition of the class was also considered, but the percentage of the class that was female was not a significant determinant in any specification.
8. The effect that the time of day that a given class meets might have on SET scores was also considered but, in no case, were these effects significant in a statistical or economic sense.
9. Kennedy and Siegfried (1997) noted that class size could also be endogenous if better teachers were assigned to larger classes, but they found no evidence of this.
10. It is possible that the relationship between class size and SET score is nonlinear: Perhaps above a certain class size further increases in the number of students is perceived in a positive light by students because of advantages of anonymity. To examine this possibility, I also included the square and cube of class size as regressors in the models presented here. In all cases, I failed to reject the hypothesis that the class size-SET score relationship is a linear one.
11. The data do not include information about semesters of experience teaching prior to an instructor joining the faculty at UNT.
12. Neither the number of classes a given instructor taught in a given semester nor whether the instructor had recently taught a particular course were significant determinants of SET scores in any specification and therefore were not included in the models.
13. In the models using principles classes, the relatively large Hausman statistics for the models examined argue for the use of the fixed-effects (FEM) rather than the random-effects (REM) framework. The *p* values for these statistics were in both cases below 0.05. However, the Hausman statistics were smaller for the models using upper-division classes, meaning that one cannot reject the hypothesis that the REM is appropriate. As it happens, results from the FEM and the REM were essentially identical, and because the use of the FEM makes adjusting instructor rankings substantially simpler, the FEM was used. REM results are available from the author upon request.
14. To explore how the differences in specification might affect the results, I applied the Isely and Singh (2005) approach to the data I used in this research. My specification assumed that the influence of instructor and the influence of course are additive and separate. In fact, the specification used here was a restricted version of that used by Isely and Singh, so an *F* test of the hypothesis that it is appropriate to treat the fixed effects as I did can be carried out. The *F* statistics were 0.41 and 0.79 for the principles and upper-division data, respectively, so the assumption that separate and additive effects are appropriate cannot be rejected. Furthermore, there were no important differences in the results, either in the magnitudes of the estimated nonfixed-effect coefficients or in their estimated standard errors. These results are available from the author upon request.
15. The percentage of the class that was majoring in economics was not included as a regressor in the principles regressions because less than 1 percent of students in these classes were economics majors.
16. In the results that follow, I present an overall constant that was recovered in the manner described by Greene (2000, 565). Each instructor-specific fixed-effect represents by how much that instructor's intercept differed from the overall constant for any particular time period. Similarly, each semester-specific fixed-effect represents the amount by which a particular semester's intercept differed from the overall constant for any particular instructor.
17. The *F* statistic was 21.43.
18. The *F* statistic was 2.35. The critical value in this case was 2.02.
19. An *F* test of the hypothesis that the course-specific fixed-effects are jointly insignificant can only be rejected at the 0.20 Type I error level (the *F* statistic was 1.28). Nevertheless, these dummies were included on the argument that it is important to control for course-specific heterogeneity.
20. With an *F* statistic of 4.65, the hypothesis that the instructor-specific effects jointly did not matter can be rejected at the 0.01 Type I error level.

21. The hypothesis that semester-specific effects jointly did not matter can be rejected at the 0.12 Type I error level (the *F* statistic equals 1.44).
22. I also examine the Isely and Singh (2005) specification that used fixed effects for each course of each instructor (see note 14). I also applied a Davidson-MacKinnon *J* test to this specification, but the conclusion was the same: The surprise variable was not clearly preferable to EXPGRADE.
23. It should be noted that the larger class sizes and lower response rates in the present data compared with that used by Isely and Singh may make it more likely that specifications that employ the surprise variable will be rejected.
24. Expected grade, which comes from the evaluation process, should be used in the adjustment rather than actual realized grade. If the latter were used, instructors could "game" the process by leading students to expect high grade but then assigning them low ones. The use of expected grade removes this possibility.
25. Semester-specific effects were not included in the adjustment proposed in equation (2), on the grounds that they affect each instructor in a given semester equally.

REFERENCES

Aigner, D. J., and F. D. Thum. 1986. On student evaluation of teaching ability. *Journal of Economic Education* 17 (Fall): 243–65.

Becker, W. E., and J. Powers. 2001. Student performance, attrition, and class size given missing student data. *Economics of Education Review* 20 (August): 377–88.

Becker, W. E., and M. Watts. 1999. How departments of economics evaluate teaching. *American Economic Review Papers and Proceedings* 89 (2): 344–49.

Danielsen, A. L., and R. A. White. 1976. Some evidence on the variables associated with student evaluations of teachers. *Journal of Economic Education* 7 (Spring): 117–19.

Davidson, R., and J. G. MacKinnon. 1981. Several tests for model specification in the presence of alternative hypotheses. *Econometrica* 49 (3): 781–93.

Dilts, D. A. 1980. A statistical interpretation of student evaluation feedback. *Journal of Economic Education* 11 (Spring): 10–15.

Greene, W. H. 2000. *Econometric analysis*. 4th ed. Upper Saddle River, NJ: Prentice-Hall.

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? *Journal of Economic Education* 36 (Winter): 29–42.

Kennedy, P. E., and J. J. Siegfried. 1997. Class size and achievement in introductory economics: Evidence from the TUCE III data. *Economics of Education Review* 16 (4): 385–94.

Krautmann, A. C., and W. Sander. 1999. Grades and student evaluations of teachers. *Economics of Education Review* 18 (1): 59–63.

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): 403–16.

Nelson, J. P., and K. A. Lynch. 1984. Grade inflation, real income, simultaneity, and teaching evaluations. *Journal of Economic Education* 15 (Winter): 21–37.

Nichols, A., and J. C. Soper. 1972. Economic man in the classroom. *Journal of Political Economy* 80 (5): 1069–73.

Seiver, D. A. 1983. Evaluations and grades: A simultaneous framework. *Journal of Economic Education* 14 (Summer): 32–38.

Siegfried, J. J., and J. Vahaly, Jr. 1975. Sample bias of unannounced student evaluations of teaching. *Journal of Economic Education* 6 (Spring): 137–39.

Tronetti, R. J., 2001. Does class size matter? Evidence from panel data estimation. Master's thesis, University of Central Florida.