Leveling the Playing Field: Should Student Evaluation Scores be Adjusted?*

Michael A. McPherson, *University of North Texas*

R. Todd Jewell, *University of North Texas*

Objectives. Colleges and universities routinely use evaluation scores to assess the quality of an instructor's teaching for purposes of promotion and tenure and for merit-raise allocations. This article attempts to identify the determinants of these scores, and to suggest ways that departments' numerical rankings of instructors might be adjusted. *Method.* This article applies a feasible generalized least squares model to a panel of data from master's-level classes. *Results.* We find that instructors can "buy" better evaluation scores by inflating students' grade expectations. Also, the teaching experience of instructors has an impact on evaluation scores, but this effect is largely seen as an increase after tenure is granted. In addition, we find evidence of a bias against nonwhite faculty. *Conclusion.* Our results suggest that an adjustment to the usual departmental rankings may be in order.

Researchers have been interested in student evaluation of teaching (SET) at the college and university level, and in their determinants, for more than a half-century. 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. There is also some evidence that SET scores affect student retention rates (Langbein and Snider, 1999). Although some strands of the literature debate whether 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. The current article contributes to the literature in several areas. First, we explore

*Direct correspondence to Michael A. McPherson, Department of Economics, PO Box 311457, University of North Texas, Denton, TX 76203-1457 ⟨mcpherson@unt.edu⟩. The authors will share all data and coding information with anyone wishing to replicate their results. The authors thank Myungsup Kim, Margie Tieslau, Janice Hauge, Young Se Kim, David Molina, Jeffrey Rous, and two anonymous referees for helpful suggestions.

SOCIAL SCIENCE QUARTERLY, Volume 88, Number 3, September 2007

the effect of instructor gender and race on SET scores. The consensus in the literature has been that these factors have little or no effect on an instructor's evaluation scores (Feldman, 1993; Krautmann and Sander, 1999; Tronetti, 2001; Campbell, Gerdes, and Steiner, 2005), although Hamermesh and Parker (2005) find evidence that minority instructors' and female instructors' ratings are lower than those of their white and male counterparts, ceteris paribus. Second, our data depart from most other literature in that our data focus on SET scores from master's-level courses in a department that places a high value on teaching. In the earlier literature, when graduate course data are used at all they are commonly pooled with undergraduate data (e.g., Seiler, Seiler, and Chiang, 1999; Mason, Steagall, and Fabritius, 1995). Given the fundamental differences in the structure and content of graduate relative to undergraduate courses, such pooling is likely to be inappropriate.

Third, we apply a feasible generalized least squares approach to a panel of data in an effort to properly account for the 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, Steagall, and Fabritius, 1995; Tronetti, 2001; Isely and Singh, 2005; McPherson, 2006). 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 one way rankings could be usefully adjusted to account for extrinsic factors that might otherwise pollute the rankings. For example, given that we find evidence that nonwhite instructors receive lower evaluation scores, adjusting SET scores may be more important when considering the relative ranking of white and nonwhite instructors. The data we use involve master's-level economics classes but, arguably, our results are more broadly applicable. In particular, economics is a quantitative discipline, especially at the graduate level. To the extent that other disciplines share this particular characteristic, our results should be generally informative.

Data and Empirical Methods

The data were obtained from the University of North Texas (UNT) Academic Records Office and from the UNT Department of Economics. The data set represents 24 consecutive semesters between January 1994 and December 2005 and comprises 280 individual graduate classes taught by a total of 22 different instructors. UNT is a comprehensive state university with more than 32,000 students. At the graduate level, UNT has both doctoral and master's programs. The Department of Economics offers only a terminal master's degree and normally services approximately 40 economics graduate students as well as master's and Ph.D. students from other departments. The department places emphasis on classroom teaching at the graduate level, having judged that this is an area in which it has a

TABLE 1

Summary Statistics

	Mean	*SD*
Evaluation	3.338	0.404
Expgrade	3.573	0.286
Pctfemale	0.471	0.229
Leveling	0.193	0.395
Theory	0.129	0.335
Econometric	0.196	0.398
Oneday	0.825	0.381
Size	25.464	11.445
White	0.800	0.401
Male	0.782	0.414
Age	41.843	8.562
Experience	19.954	12.882
Adjunct	0.054	0.226
Tenure	0.554	0.498
Evalnumber	8.518	6.145

$N = 280$.

competitive advantage in attracting students over Ph.D. programs and other master's programs in the region. The UNT program is similar to other terminal master's programs in economics at other large state universities; thus, our data set is representative of that group. Furthermore, it is likely that our results will generalize to other disciplines that share economics' quantitative emphasis. The variables used in this study are discussed below, and summary statistics are given in Table 1.

The Department of Economics uses an SET instrument that includes 20 statements with which students are asked to gauge their level of agreement: agree strongly; agree moderately; disagree moderately; or disagree strongly, with some statements made in a positive manner and some in a negative manner. The measure of SET scores used in the present study (EVALUATION) is an average of the responses of all students to the following statements: "I would take another course that was taught this way"; "The instructor did NOT synthesize, integrate, or summarize effectively"; "Some things were NOT explained very well"; "I think the course was taught quite well"; and "Overall, the course was good." EVALUATION ranges from 1 (indicating strong disagreement with positive statements and strong agreement with negative statements on average) to 4 (strong agreement with positive statements and strong disagreement with negative statements on average), with 4 representing the best possible SET score.[1]

[1]We evaluate the responses to each of the questions using both Cronbach's alpha and principal components analysis. For these five questions, Cronbach's alpha is relatively high

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 expected grade (EXPGRADE) and the proportion that is female (PCTFEMALE).[2] EXPGRADE is measured on a four-point scale, averaging 3.57 for these data; a priori, one would expect higher evaluation scores to be correlated with higher expected course grades.[3] The gender composition of the course may influence SET scores if evaluation standards vary by the gender of the student. We also include semester time dummies to control for changes in the composition and preferences of graduate students over time. This time trend may also pick up changes in the composition of instructors over time.[4]

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: LEVELING = 1 if the course is one of four courses for incoming graduate students who do not have the proper educational background; ECONOMETRIC = 1 if the course is one of the four econometric courses offered to master's students; and THEORY = 1 if the course is part of the sequence of four microeconomic and macroeconomic theory courses. The excluded category comprises courses that are electives for most graduate students.[5] Another important course-specific characteristic is the number of days per week that the course meets. This aspect is modeled with a dummy variable equal to 1 if the course meets once a week (ONEDAY). 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, 82.5 percent of graduate courses meet once a week for three hours. The remainder meet three times per week for one hour (2 percent) or twice per week for 1.5 hours (15.5 percent). A final course characteristic is class size. Some contributions to the literature have found an inverse relationship between class size and SET score; normally, larger classes are found to have lower average SET scores. For example, Isely and Singh (2005) and Tronetti (2001) find such an effect using data from undergraduate classes; Boex (2000) reports similar findings for graduate classes.

(0.89), the first principal component accounts for 70 percent of the variance among the five questions, and the factor loadings are approximately equal. Thus, we conclude that an SET measure computed as a simple average of these five questions is reliable and valid. SET forms are distributed without announcement beforehand at the end of the semester and are anonymous. Eighty-two percent of enrolled students completed the evaluation questionnaire.

[2]EXPGRADE represents the average expected grade of students who fill out the evaluation form. PCTFEMALE is based on the entire class.

[3]Research on undergraduate SET scores indicates that expected grade may be endogenous (e.g., Seiver, 1983; Nelson and Lynch, 1984). Employing a Hausman test (results available on request), we find no evidence of endogeneity in our sample of graduate courses.

[4]The semester dummies are not reported for the sake of brevity (results available on request).

[5]The number of courses (25) makes inclusion of course-specific dummy variables impractical. In addition, the course dummies and instructor-specific effects are highly collinear.

The third group of SET score determinants comprises 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. A specification test indicates that a panel data model using either random or fixed instructor-specific effects is appropriate.[6] We choose a random-effects model since it allows for the inclusion of time-invariant regressors, such as gender and race, and because it is more efficient than a fixed-effects model.

Observable characteristics of instructors include the instructor's gender (MALE) and race (WHITE), total semesters of university teaching EXPERIENCE (and its square), AGE (and its square), whether the instructor is an ADJUNCT, and whether the instructor has been granted TENURE. 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, some research exists that suggests that students perceive female instructors differently from their colleagues who are men. Basow (1998) and Andersen and Miller (1997) argue that male and female instructors tend to approach teaching differently and that, in addition, students have differential expectations of how male and female instructors should behave, and therefore would react to instructors differently according to gender in the evaluation process. Basow and Silberg (1987) find statistical evidence that male and female students each tend to give female instructors less favorable SET scores. SET scores are expected to increase with EXPERIENCE, since more time in the classroom should increase the quality of one's teaching. EXPERIENCE includes previous teaching experience at all universities, not just at UNT. An instructor who is an ADJUNCT is not on tenure track and has no research and limited service responsibilities. Thus, an ADJUNCT is hired for one purpose: teaching. We expect that such instructors will have higher SET scores, all else equal. The possible effect of TENURE on SET scores is unclear a priori. Upon the awarding of TENURE, some instructors may decrease effort put into teaching, while others may feel liberated from research and increase teaching effort.[7]

[6]A Hausman test indicates that the assumption of the random-effects model concerning the orthogonality of the random effects and the regressors is appropriate. The chi-square statistic (14 degrees of freedom) is 11.36, which is insignificant at any conventional level. Thus, we cannot reject the null of no correlation between the random effects and the regressors.

[7]We attempted to include other measures in the analysis: the evaluation response rate; whether the course was required for a student's degree; and the time of day the course was taught. None of these variables was significant (results available on request).

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 the following equation:

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

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

Results

The equation presented above is estimated using feasible generalized least squares (FGLS), and the results are reported in Table 2. We use a weighted estimation since the error variances of SET scores will be larger for courses with fewer students. 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 $\sigma^2{}_u$. FGLS allows the variance of u to vary across instructors, which is important in our case since we have heteroskedasticity due to unbalanced panels; the average instructor in our sample has taught 13 courses, with the number of courses taught ranging from as few as two to as many as 46.

The results in Table 2 suggest that expected grade significantly affects SET scores, implying that instructors can induce higher scores by increasing the grade expectations of their graduate students. Some researchers have found evidence that SET scores can be "bought" in this manner in undergraduate classes (Aigner and Thum, 1986; Greenwald and Gillmore, 1997; McPherson, 2006). The coefficient on EXPGRADE implies that a one point increase in expected grade will lead to approximately a 0.25 increase in SET score. To put this result in context, assume an instructor inflates grade expectations such that students in her class go from expecting the mean (3.573) to expecting a 4.0. In this case, the instructor's SET score is predicted to increase by 0.11 points. Classes comprising larger proportions of female students tend to give neither higher nor lower scores to their instructors.[8]

[8]Some research suggests that SET scores will be influenced by the relationship between the gender of the instructor and that of the students. For instance, Bachen, McLoughlin, and Garcia (1999) present evidence that female students have a tendency to rate female professors more favorably, while male students do not seem to judge their instructors differently according to gender, and Basow (1998) reports evidence that students tend to assign more favorable evaluations to instructors of the same gender as the student. Another model (results

TABLE 2

FGLS Estimates

Constant	4.6105***
	(0.2272)
Expgrade	0.2544***
	(0.0711)
Pctfemale	0.0265
	(0.0890)
Leveling	−0.1232***
	(0.0489)
Theory	−0.1816***
	(0.0575)
Econometric	0.1881***
	(0.0764)
Oneday	−0.2035***
	(0.0619)
Size	−0.0112*
	(0.0065)
$Size^2$	0.0002*
	(0.0001)
White	0.0829**
	(0.0435)
Male	−0.0187
	(0.0698)
Age	−0.0831***
	(0.0250)
Age^2	0.0007***
	(0.0003)
Experience	0.0055
	(0.0060)
$Experience^2$	−0.0001*
	(0.0001)
Adjunct	0.2476***
	(0.0730)
Tenure	0.1380***
	(0.0527)
Pseudo-R^2	0.5581

*Significant at 10% level; **significant at 5% level; ***significant at 1% level.

NOTE: Dependent variable = EVALUATION (weight = EVALNUMBER). Standard errors in parentheses. $N = 280$.

The course-specific measures are generally significant determinants of SET scores. Leveling and theory courses each have lower scores than elective courses, while scores from econometric courses are significantly higher than elective courses. Instructors of courses that meet one day a week have 0.20

available on request) was estimated with an interaction term between PCTFEMALE and MALE, but the estimated coefficient was insignificant.

lower evaluation scores, possibly implying that students would rather take courses that meet multiple times per week. Finally, we find that SET scores decrease with class size up to 28 students and increase thereafter.[9]

Turning to the instructor-specific measures, Table 2 suggests that SET scores are affected by observable characteristics. Race appears to play a significant role in SET scores in our data, with white instructors earning 0.08 higher SET scores than their nonwhite colleagues.[10] Instructor gender does not play a significant role in determining SET scores in our sample. If we assume that race plays no role in teaching quality or ability, this outcome may indicate that students evaluate teaching differently based on the instructor's race. As expected, adjunct faculty receive significantly better scores than tenure-track faculty.

We are interested in the impact of academic experience on SET scores, embodied in the instructor-specific measures EXPERIENCE and TENURE. Although only EXPERIENCE^2 is marginally significant individually, EXPERIENCE and EXPERIENCE^2 are jointly significant at the 5 percent level. Somewhat surprisingly, the coefficients indicate that another semester of teaching experience increases EVALUATION up to 20 semesters (i.e., the sample mean) and decreases SET scores after that. Less surprising is the coefficient on TENURE; the results indicate that a nontenured faculty member is expected to receive lower SET scores than a tenured faculty member. Assuming a six-year probationary period and taking the average SET score as a benchmark for Year 1 (3.338), an instructor's SET score at the end of the sixth year is predicted to be 3.404, and it is predicted to increase to 3.542, 6 percent above the sample average, after earning tenure. What is it about the granting of tenure that leads to higher SET scores? Numerous possibilities exist, but two seem especially worthy of note. First, earning tenure at UNT is a function of both research and teaching, with possibly a heavier emphasis on teaching than at comparable schools. Tenure status and SET scores may be positively correlated simply due to the fact that the worst teachers are more likely to be denied tenure. Second, the granting of tenure may be associated with a lessening of time spent on research, thus freeing up more time for teaching.

We can also examine the effects of instructor age on SET score, independent of teaching experience. The results in Table 2 demonstrate that instructors receive lower SET scores as they age up to 55 years and higher

[9]Some courses are offered to both graduates and undergraduates simultaneously. SIZE includes all students, both undergraduate and graduate, and EVALNUMBER is the number of evaluation forms filled out by graduate students.

[10]This conclusion should be viewed with some skepticism as white instructors teach the majority of courses in our sample and 19 of the 22 instructors are white. The nonwhite category includes Hispanics, blacks, and Asians. In our sample, there are 224 observations for white instructors (80 percent), while 53 are Hispanic (19 percent), and three are Asian (1 percent). In addition, since we do not have data on the racial composition of the students, we cannot test for an interactive effect between the race of the instructors and students.

scores at subsequent ages, although the magnitude of the effect is small. It is interesting that, after holding teaching experience constant, students appear to reward youthfulness in their instructors. One explanation for this finding may be a correlation between an instructor's age and his or her perceived "attractiveness," at least in terms of how students judge this quality; since we do not include a measure of student-perceived attractiveness, the variable AGE may be picking up this effect. Two recent studies find contradictory results with respect to instructor attractiveness. Hamermesh and Parker (2005) that better-looking instructors receive higher SETs than their "less attractive" colleagues. Campbell, Gerdes, and Steiner (2005) explore the effect of instructor beauty on SET score, but find no significant relationship.

Adjusted Rankings

Several researchers (Danielsen and White, 1976; Mason, Steagall, and Fabritius, 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 race). For instance, a department could produce a ranking based on instructor-specific random effects, which would hold constant all observable effects.[11] The random effects might be thought of as overall or longer-term indicators of instructors' teaching, taking out the effects of race, gender, teaching experience, and all other observable effects. Although 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, if it were the case that a large part of an instructor's random effect was the result of factors such as personality or appearance, then an adjusted ranking based on random effects might not "improve" the rankings in any meaningful way.

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 assigns all instructors the same race. A ranking based on this measure would effectively remove the disadvantages under which nonwhite instructors may be operating, while also controlling for differences in other observable factors.

[11]The FGLS random-effects estimator does not produce an estimate of the individual-specific effect. Following Greene (2003:296), we could use the mean of the differences between EVALUATION and its predicted value as an estimate of the instructor-specific random effect.

TABLE 3

Adjusting Semester Rankings, Overall Average

Instructor	Race	Untenured Obs.	Tenured Obs.	Expected Grade	Raw Ranking	Adjusted Ranking*
A	white	3	0	3.533	1	7
B	white	20	26	3.592	2	2
C	white	11	13	3.758	3	3
D	white	13	0	3.506	4	6
E	white	0	14	3.645	5	12
F	white	7	6	3.573	6	8
G	white	10	0	3.504	7	10
H	nonwhite	3	0	3.444	8	9
I	nonwhite	19	0	3.407	9	4
J	white	6	0	3.647	10	1
K	white	2	8	3.310	11	5
L	white	0	34	3.615	12	14
M	nonwhite	0	34	3.604	13	15
N	white	9	0	3.548	14	13
O	white	2	0	3.713	15	11
P	white	5	13	3.540	16	16
Q	white	0	7	3.313	17	17

*White = 0, tenure = 0, and expgrade = 3.573.

In Table 3, we present a ranking based on a specific adjustment for tenure-track faculty over the entire 24-semester time period.[12] The adjustment treats all instructors as nonwhite and without tenure, effectively removing the advantage that being white and tenured seems to confer on some instructors. Furthermore, the ranking adjusts for differences in expected grade, effectively removing the impact of instructor behavior designed to influence SET scores, by assigning each instructor the mean expected grade of the sample. The adjusted ranking shows that if all instructors were of the same race and tenure status and if all classes had the same expected grade, there would be changes in the relative ranking of instructors; for some, the change in ranking is rather dramatic. Consider Instructor J, a nontenured instructor. She would be ranked 10th out of 17 tenure-track instructors using the raw SET scores, but she would be the top-rated instructor were the playing field to be level with respect to race, tenure status, and expected grade. Instructors I and K also increase their rankings with the adjustments, and the data suggest that these improvements are related to the instructor's relatively low expected grades. In most cases, we would expect to see tenured faculty with high expected grades to do comparatively worse as a result of this adjustment; Instructor E illustrates this effect.

[12]We exclude adjunct instructors from the adjusted rankings.

TABLE 4

Adjusting Semester Rankings, Spring 2003

Instructor	Race	Tenure	Expected Grade	Raw Ranking	Adjusted Ranking*
M	nonwhite	yes	3.500	1	6
B	white	yes	3.622	2	2
C	white	yes	3.813	3	3
F	white	yes	3.429	4	4
P	white	yes	3.333	5	11
I	nonwhite	no	3.560	6	7
E	white	yes	3.600	7	9
L	white	yes	3.702	8	10
J	white	no	3.700	9	1
N	white	no	3.429	10	5
K	white	yes	2.500	11	8

*White = 0, tenure = 0, and expgrade = 3.573.

Table 3 provides adjusted rankings for the entire sample period. In some cases, departments will not use long-term averages when analyzing SET scores; instead, they may rely on semester-by-semester rankings. The procedure outlined above can easily be used to adjust the rankings each semester. To illustrate, Table 4 presents the raw and adjusted ranking for the spring 2003 semester, once again removing the effects of race, tenure status, and expected grade.[13] Similar to the overall ranking of Table 3, the semester-specific ranking of Table 4 shows that the ranking of some instructors in the UNT Economics Department improves dramatically when SET scores are adjusted. As is the case in Table 3, Instructor J becomes the top-rated instructor after the adjustment.

Of course, it is debatable whether adjustments for race, tenure status, or expected grade are appropriate. Although not reported here, one could also compute predicted SET scores adjusting for whatever variables are thought to pollute the rankings. For example, our results indicate that students give relatively higher scores in econometric courses. Adjustment of SET scores to eliminate this effect on instructor rankings is one way to minimize the disincentive for some faculty members to teach courses that students like less than others. There are certainly other factors that are beyond an instructor's control that apparently have an effect on SET scores, and these factors will vary by university, department, and student level. Ranking adjustment would permit departments to take such factors into account as well when considering a ranking of instructors based on SET scores.

[13]We report the rankings based on predicted values of evaluation for a single, representative semester. Rankings for all semesters show a similar pattern (results available on request).

Conclusion

The issue of whether the SET process actually measures quality of teaching output (and therefore whether SETs should be used to evaluate instructors) may never be settled conclusively, but the fact remains that SETs are important components of both the promotion and tenure and merit-raise allocation processes at many U.S. universities. As such, a better understanding of the factors that drive evaluation scores is a worthwhile goal. This article uses a feasible generalized least squares approach to examine a panel of data comprising 280 individual master's-level classes over 24 consecutive semesters. This approach represents one of the few attempts to properly account for individual-specific unobservable effects. It is also unusual in its focus on graduate-course data from a university that values teaching.

Several principal findings emerge. First, we find evidence that instructors can increase SET scores in graduate courses by inflating grade expectations. In addition, certain other factors specific to courses (and largely out of the control of individual instructors) influence an instructor's evaluation score. For example, teaching theory courses and teaching once weekly in a three-hour lecture setting significantly worsens instructors' SET scores within our sample. In a real sense, ranking instructors by their average evaluation scores may reward those who are lucky enough to be selected to teach elective courses that meet multiple times per week. Our results suggest that it may be important to explore the adjustment of rankings to eliminate the effects of such factors. We show that there are clear differences between raw and adjusted rankings and that these differences are substantial for some instructors.

We also find evidence that nonwhite instructors are at a disadvantage in terms of SET scores relative to their white colleagues. That is, controlling for all other observable effects of courses, instructors, and semesters, there remains a gap of approximately 0.08 points (on a four-point scale) between the scores of white and those of nonwhite instructors. If race has no correlation with teaching ability, then, arguably, this is a disadvantage that departments ought to correct by adjusting evaluation scores. Similarly, our results indicate that on receipt of tenure an instructor's SET score improves by nearly 0.14 points, ceteris paribus. This difference exists even after controlling for the effects of instructor age and experience, so evidently there is some discrete change at or near the time of tenure that leads to better student evaluations. Once again, while debatable, it is possible to argue that this constitutes a disadvantage under which junior faculty may be operating. If a particular academic unit deems this to be the case, an adjustment to the rankings such that all instructors are treated as having the same tenure status might be appropriate.

As our results show, such adjustments to the rankings can be significant in certain individuals' cases. The particular examples of adjustments we

explored (race, tenure status, and expected grade) are only illustrative of the broader point. There may be other adjustments that might be important to some departments. For example, our results seem to indicate that evaluation scores decrease as a faculty member ages. Should younger faculty be effectively stripped of this advantage by means of a ranking adjustment? The issue of bias in SET scores is one that each department should discuss, and one that each department may resolve in its own fashion. In general, our results indicate that academic units may find it useful to explore possible biases in the rankings; furthermore, departments may find that adjustments based on some set of criteria are a valuable exercise.

REFERENCES

Aigner, D. J., and F. D. Thum. 1986. "On Student Evaluation of Teaching Ability." *Journal of Economic Education* 17:243–65.

Andersen, K., and E. D. Miller. 1997. "Gender and Student Evaluations of Teaching." *PS: Political Science and Politics* 30:216–19.

Bachen, C. M., M. M. McLoughlin, and S. S. Garcia. 1999. "Assessing the Role of Gender in College Students' Evaluations of Faculty." *Communication Education* 48:193–210.

Basow, S. A. 1998. "Student Evaluations: The Role of Gender Bias and Teaching Styles." In L. H. Collins, J. C. Chrisler, and K. Quina, eds., *Career Strategies for Women in Academe: Arming Athena*. Thousand Oaks, CA: Sage.

Basow, S. A., and N. T. Silberg. 1987. "Student Evaluations of College Professors: Are Female and Male Professors Rated Differently?" *Journal of Educational Psychology* 79:308–14.

Boex, L. F. J. 2000. "Attributes of Effective Economics Instructors: An Analysis of Student Evaluations." *Journal of Economic Education* 31:211–27.

Campbell, H. E., K. Gerdes, and S. Steiner. 2005. "What's Looks Got to Do with it? Instructor Appearance and Student Evaluations of Teaching." *Journal of Policy Analysis and Management* 24:611–20.

Danielsen, A. L., and R. A. White. 1976. "Some Evidence on the Variables Associated with Student Evaluations of Teachers." *Journal of Economic Education* 7:117–19.

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* 34:151–211.

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

Greenwald, A. G., and G. M. Gillmore. 1997. "Grading Leniency is a Removable Contaminant of Student Ratings." *American Psychologist* 52:1209–17.

Hamermesh, D. S., and A. Parker. 2005. "Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity." *Economics of Education Review* 24:369–76.

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

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

Langbein, L. I., and K. S. Snider. 1999. "The Impact of Teaching on Retention: Some Quantitative Evidence." *Social Science Quarterly* 80:457–72.

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* 14:403–16.

McPherson, M. A. 2006. "Determinants of How Students Evaluate Teachers." *Journal of Economic Education* 37:3–20.

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

Seiler, M. J., V. L. Seiler, and D. Chiang. 1999. "Professor, Student, and Course Attributes that Contribute to Successful Teaching Evaluations." *Financial Practice and Education* 9:91–99.

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

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