Evaluating College Teaching Performance: The Case of Principles of Economics Classes

Joachim Zietz, Mark L. Wilson, and Howard H. Cochran, Jr.*

Abstract

The paper suggests a methodology to remove potential bias from performance evaluations, which may arise, for example, if those being evaluated cannot control key factors that affect their performance score. The methodology is illustrated for the case of instructors' teaching evaluations in principles of economics classes at a large comprehensive university. Standard student evaluations and the results of an end-of-term test of student knowledge are adjusted for student and class characteristics using simple regression techniques. In addition, student evaluations are adjusted for grade inflation. The results of the suggested methodology are compared to those derived from the common evaluation method based on unadjusted student evaluations.

Introduction

Regular performance evaluation has become accepted practice in both for-profit and not-for-profit organizations. Performance measurement in the non-profit sector, including college education, may appear at first sight to be harder than measuring performance in the for-profit sector of the economy because one cannot rely on profit as the ultimate measure of success or failure. But performance evaluation is not without problem in the for-profit sector either.

Consider for example the issue to what extent a declining profit level can be attributed to the substandard performance of a particular person or organization. If the industry or the whole economy is moving into a downturn and profit levels decline throughout, it would little sense to put all the blame for declining profits on the person or organization to be evaluated. In fact, profits may have declined significantly more had it not been for the exceptional performance of those to be evaluated. Not taking into account such extenuating circumstances may have the undesirable consequence that one could end up losing some of the best people. This point seems sometimes lost in public discussions when corporate managers receive pay increases even though the company's profits are declining. This simple example points to the fact that, ideally, performance evaluation should be limited to that part of the outcome measure that can be controlled by the person or organization being evaluated. That basic principle should apply not only when the evaluation relies on an objective outcome measure, such as profit or cost, but also when the evaluation relies on perceptions and opinions, such as when employees evaluate the performance of their superior or manager or when students evaluate their teacher.

The purpose of this paper is to suggest some practical ways to avoid typical problems of performance evaluations and, thereby, to make evaluations more meaningful and also more acceptable to those being evaluated. The suggested methodology is demonstrated for the classroom performance evaluation of college

*Joachim Zietz is Professor of Economics, Middle Tennessee State University, Murfreesboro, Tennessee, E-mail: jzietz@mtsu.edu. Mark Wilson is Research Director, West Virginia Insurance Commission, Charleston, West Virginia, E-mail: Mark.Wilson@wvinsurance.-gov. Howard Cochran is Professor of Economics & Management, Belmont University, Nashville, Tennessee, E-mail: cochranh-@mail.belmont.edu. Financial support for the first author by the Faculty Research & Creative Activity Committee at Middle Tennessee State University is gratefully acknowledged.

teachers. The particular example used in this paper relates to the evaluation of principles of economics classes taught at a large comprehensive university.

The paper is organized as follows. The next section discusses the suggested methodology and the data that are employed to illustrate the methodology. Subsequently, the performance evaluation measures are constructed and interpreted. The paper ends with a brief discussion of its main results.

Methodology and Data

The Adjusted Performance Measures

The suggested adjustments to typical performance measures will focus on evaluating the classroom performance of college teachers. However, as will become apparent, many of the incentive problems that arise in the evaluation of college teaching have clear counterparts in other areas where the suggested methodology could be applied, for example in the evaluation of sales personnel.

The evaluation of classroom performance of teachers at universities relies to a large extent on student evaluations. Typical student evaluations have a number of known problems. First, good teaching evaluations do not necessarily mean that students learn a lot. In fact, there is credible evidence of a low or even negative correlation between the two (e.g., Abrami et al. 1990, Gramlich and Greenlee 1993). Second, student evaluations are likely influenced by factors that cannot be attributed to instructors, such as time-of-day of class meetings (Mirus 1973) or student ability (Seiver 1982, Nelson and Lynch 1984). Third, instructors may "buy" good student evaluations with grade inflation (Dilts and Fatemi 1982, Boex 2000). All of these and other problems give student evaluations rather limited credibility among faculty (Simpson and Siguaw 2000).

To address the first problem, it is suggested that both student evaluations of teaching (SETs) and an objective measure of student learning (EXAM) are needed to properly assess teaching performance. A weighted average of the two measures may be a simple solution. To cope with the second and third problems, both measures are adjusted by a regression-based method.[1] The regression-based method relies on two separate regression equations. One equation is estimated for EXAM scores, the other for SET scores. Student and class characteristics that are not under the control of the instructor serve as the predictor variables in these two regressions.

Following estimation of the two regressions, the observed score $y_{i,j}$ for teacher j ($j = 1,\ldots, m$) and student i ($i = 1,\ldots, n$) is compared to the one that is predicted by the respective regression equation. The predicted score is identified as $\hat{y}_{i,j}$. Following common practice in statistics and human resource management (Cascio 1998), only deviations of actual from predicted scores that are larger than one standard error of the estimate of the underlying regression ($\hat{\sigma}$) count as being significantly different from what can be expected.

Two alternative performance measures are proposed on the basis of the underlying regressions. The first measure is the percentage of students with significantly better than expected scores minus the percentage of students with significantly worse than expected scores. The expected scores are the ones predicted by the underlying regression equations plus or minus one standard error of the estimate of the underlying regression. More formally, for each instructor j that is being evaluated, the first measure (M_1) is given as

$$M_1 = \frac{1}{n}\sum_{i=1}^{n}\left[I_{i,j}\left(y_{i,j} - \hat{y}_{i,j} > \hat{\sigma}\right) - I_{i,j}\left(y_{i,j} - \hat{y}_{i,j} < -\hat{\sigma}\right)\right], \quad \forall j$$

where $I_{i,j}$ is a 0/1 indicator variable that is unity if the condition in parenthesis immediately following the indicator variable is fulfilled and zero otherwise. Hence, in the extreme case where each student i of professor j

[1] Some previous efforts have been made to calibrate teaching performance measures. For example, Rose (1975) observed that SET scores reflect class conditions over which the instructor has no control. Dilts (1980) and Zangenehzadeh (1988) adjust SET scores for grade inflation. Shmanske (1988) ranks faculty on the basis of SET scores and cognitive outputs when class performance of a student in a subsequent class is predicted by the instructor taken for the first class. Mason et al. (1995) re-rank faculty after adjusting SET scores for influences that the instructor cannot control. Bailey et al. (2000) find at least three essential control variables are needed to validate SETs.

has a score that is within plus or minus one standard error of the estimate, M_1 has a value of zero. In the extreme case, where half of the students score outside and above the one standard error band around the regression line while half of them score outside and below the error band, M_1 also has a value of zero. Since M_1 is derived by summing the *number* of occurrences of unusually high scores and subtracting the *number* of unusually low scores, it does not matter whether the scores are very far from the one standard error band around the regression or not. By contrast, the distance from the error band matters in the measure (M_2) that is introduced next.

If one wants to make sure that score size matters for evaluation purposes, then the measure M_2 would be relevant. M_2 is based on the difference between actual score and estimated error band. This difference is calculated for each student. Then, the negative differences, as derived from students with lower than predicted scores, are summed and subtracted from the sum of the positive differences, as derived from students with above average scores. The total is divided by the number of students taught by professor *j*. More formally, for each instructor *j* that is being evaluated, M_2 is given as

$$M_2 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_{i,j} - \hat{y}_{i,j} - \hat{\sigma}\right) I_{i,j}\left(y_{i,j} - \hat{y}_{i,j} > \hat{\sigma}\right) - \left(y_{i,j} - \hat{y}_{i,j} + \hat{\sigma}\right) I_{i,j}\left(y_{i,j} - \hat{y}_{i,j} < -\hat{\sigma}\right)\right], \quad \forall j$$

In the extreme case mentioned earlier, where half of the students score more than a standard error above the regression line while half score more than a standard error below the regression line, M_2 can be either positive or negative. Only by coincidence would M_2 be zero as is M_1. Clearly, M_2 is a more comprehensive measure since it accounts for the different score sizes rather than just the sign of the scores. On the other hand, M_1 is less affected by extreme outliers.

Both measures, M_1 and M_2, make use of predicted scores, $\hat{y}_{i,j}$, and the predicted error of the estimate of the underlying regression, $\hat{\sigma}$. Both $\hat{y}_{i,j}$ and $\hat{\sigma}$ are based on a regression on all *m* times *n* observations.[2] The predictions are conditional on variables that are not under the control of the instructor, such as student ability. The regression explaining the EXAM score may have a somewhat different set of variables than the regression explaining the SET score. One key difference suggested here is that the regression equation for the SET score contains a variable that controls for instructor grading. The purpose of adding such a variable is to reduce the chance that instructors can "buy" good evaluations with good grades.[3] Two alternative measures of grading are tried, a student's expected grade and a student's actual end-of-term course grade. The first of these can only be obtained through a student questionnaire, while the latter can be retrieved from the university's records office.

Any regression-based approach to adjusting EXAM and SET scores faces the problem of selecting the number and type of control variables. It is unlikely that there is one correct or unique set of variables. It is likely that there will be differences of opinion across colleges within a university and across universities. This is not a weakness of the suggested adjustment method but rather its strength: it can easily adapt to different environments and preferences. As a simple example, consider the issue whether to adjust SET scores for the fact that an instructor's mother tongue is not English. Some colleges may decide to include a dummy variable for instructors with a mother tongue other than English, and, hence, to give these instructors an implicit bonus assuming that the regression equation returns a negative coefficient. Others may elect not to do that. For the purpose of the example calculations provided in this paper, no teacher characteristics are included in the regressions for either the SET or the EXAM scores.

After running the SET and EXAM regressions, M_1 and/or M_2 are calculated separately for the SET and the EXAM scores. A comprehensive measure of teaching effectiveness can then be derived by making use of both the SET and EXAM scores in their M_1 or their M_2 form. Again, the rationale for making use of the scores for both student evaluations and student performance is to avoid incentive problems, where instructors emphasize popularity with students as opposed to content and student learning.

How to combine the scores is an issue that needs to be decided by the organizational unit that is employing

[2] In practical applications of the methodology in a university environment, one could run a regression for each college or have one regression for the university as a whole.

[3] It is assumed that, on average, the prospect of good course grades will induce students to raise their student teaching evaluations (Boex 2000).

the performance measure. For example, if one employs the M_1 measure, the overall performance of instructor j could be derived as a weighted average,

$$P_j = \alpha M_{1,j}^{Exam} + (1-\alpha) M_{1,j}^{Set}$$

where α is the weight for the adjusted EXAM score and $(1-\alpha)$ the weight of the adjusted SET score. The size of weight α is a crucial decision that reflects on the mission of the unit to be evaluated. If the emphasis is mainly on student learning, then α should be selected close to or equal to unity. If by contrast there is considerable emphasis on students feeling good about the classes they attend, for example, because of student retention goals, then a lower value of α needs to be selected. If one does not want to implement a formalized approach such as represented by the P measure, one could simply opt for selecting the top performers in both the SET and EXAM categories. Those instructors that are among the top performers in both categories could then be identified as exemplary.

An issue of importance in practical applications is what use a performance measure such as P_j or its more informal equivalent is put to. Although a thorough discussion is beyond the scope of this paper, it is apparent that the organizational behavior literature (e.g., Kreitner and Kinicki, 2000) argues against rewarding some instructors for their high scores while punishing others for their low scores. Using the proposed performance evaluation in this manner is likely to result in a hostile work environment. Instead, the emphasis may have to be on rewards for the outstanding instructors. These rewards could be used to encourage others to experiment with and adopt effective techniques of instruction that raise scores. In this way, the average level of teaching effectiveness of the unit under consideration could be increased as more effective techniques become more widely adopted by faculty.

The Data to Illustrate the Methodology

The data to illustrate the suggested methodology come from more than twenty sections of principles of economics classes that were taught by eleven different instructors at a large comprehensive state university during the fall semester.[4]

The data are drawn from three major sources: (a) a questionnaire that elicits a teaching evaluation on a one to five scale as well as several biographical questions from each student,[5] (b) an objective end-of-semester achievement test, and (c) administrative records of students.

For the SET questionnaire, students were assured of the anonymity of their responses.[6] The survey yielded 571 responses from a total of approximately 800 enrolled students. The discrepancy is due to three factors: (a) some of the students dropped, (b) some were randomly absent because the questionnaire was not announced ahead of time; and (c) some provided unusable responses.

The EXAM scores come from achievement tests that were administered during the final exam period as scheduled by the university.[7] The exam consisted of 30 multiple-choice questions that were prepared by the authors.[8] A serious attempt was made to ensure that the questions were not known to students prior to exam time.

The variables used for the regression equations are detailed in Table 1. Student characteristics include (a) measures of a student's stock of human capital, such as GPA and ACT[9] scores, (b) student characteristics that

[4] The classes are about equally divided between principles of microeconomics and principles of macroeconomics.

[5] This technique differs from most research on student evaluations which use class average SET scores rather than a unique SET score for each student in the sample.

[6] Hansen and Kelley (1973) have shown that students answer differently when the data are collected for merit and promotional purposes as opposed to teacher improvement purposes.

[7] No pre-test was administered, although this would have been preferred. Hence, it was not possible to construct a measure of learning as the difference between the scores on the post-test and the pre-test.

[8] The test preparer was not teaching any of the classes that were part of this experiment. Separate tests were constructed for the principles of microeconomics and the principles of macroeconomics classes.

[9] The administrative records of some of the students in this study reported the Scholastic Aptitude Test (SAT) as a measure of aptitude. These scores were converted to an ACT score by the commonly-used conversion factor (Marco et al. 1993).

cannot be scaled up or down by a student at will, such as age or whether a class was the student's first choice, and (c) student characteristics that can be varied by each student, such as attempted class hours.

TABLE 1. VARIABLE DEFINITIONS AND SUMMARY STATISTICS

Variable	Variable Definition	Mean	Std.Dev.
Dependent Variables			
SET	Student Evaluation of Teacher; 1 to 5 scale; 1 being the minimum.	3.74	0.93
EXAM	Percentage Score on Achievement Test	56.80	14.23
Independent Variables			
Student			
GPA	Grade point average	2.63	0.69
ACT	Aptitude test score; actual ACT score or equivalent score converted from SAT score	21.01	3.80
AGE	Student's self-reported age in years	21.82	4.89
MALE	Male = 1; female = 0	0.54	0.50
ATMHRS	Attempted hours during semester	14.20	3.14
REQ	Required class = 1; elective Class = 0	0.79	0.41
JOB	Student holds job = 1; Other = 0	0.76	0.43
HRSWORK	Number of hours student works on job per week	19.67	14.71
CHOICE	First choice of class = 1; Other = 0	0.87	0.33
EXPECTED	Student's expected course grade at time of SET evaluation	2.61	0.92
GRADE	Student's end-of-term course grade	2.43	1.02
Class			
NIGHT	Night Class = 1; Other = 0	0.07	0.26

Notes: Averages and standard deviations are calculated for the maximum number of observations for each variable.

The previous literature provides the rationale for selecting the variables given in Table 1. As regards the equation predicting EXAM scores, GPA and ACT scores as well as student gender and age have been cited as strong predictors of cognitive classroom success. A positive influence of grade point average has been identified, inter alia, by Tuckman (1975), Marlin and Niss (1980), Brasfield et al. (1993), and Lopus and Maxwell (1994, 1995), and a positive influence of the ACT score, among others, by Weidenaar and Dodson (1972), Brasfield et al. (1993), and Lopus and Maxwell 1994. Typically, the variables MALE and AGE have a significant impact on EXAM scores.[10] Non-school work (HRSWORK) has produced conflicting results in the literature (Lillydahl 1990, Lopus and Maxwell 1994). Finally, a variable (CHOICE) is included in the regressions to account for evidence that students vote with their feet in their selection of classes (Leventhal et al. 1975). The CHOICE variable indicates that the student got his/her first choice of class.

Several student input variables have been identified in the literature as strong predictors of the SET

[10]See Watts and Lynch (1989), Tuckman (1975), and Lopus and Maxwell (1994) on the gender issue and Weidenaar and Dodson (1972) and Marlin and Niss (1980) on the age of students.

relationship. Specifically, students with lower grades tend to award higher teacher evaluations (Seiver 1983), although the better a student performs in a *given* class, the higher the SET tends to be (Marlin and Niss 1980). Moreover, Dilts and Fatemi (1982) have shown that required courses tend to receive lower evaluations. Age and gender seem to matter, too. Seiver (1983) finds a positive association between the students' age and the teacher's evaluation. Student commitment to the class may also be important. Lopus and Maxwell (1994) observe a negative, although insignificant, relationship between SETs and the number of hours worked. Finally, the class choice phenomenon may matter for SET scores. It is reasonable to believe that students exercise some choice in their selection of instructors, and one may suspect that they will tend to give first-choice classes higher evaluations.

Estimation Results and Interpretation

Regression Estimates Underlying the Performance Measures

The equations that are used to predict EXAM and SET scores are reported in Table 2. The first two equations report the results for the EXAM equation and the final four columns summarize the results for the SET equation. The EXAM equations do not have any apparent statistical problem. This is different for the SET equations. All of the SET equations suffer from non-normal errors, one from heteroskedasticity and one has a functional form problem.[11] As a consequence, one would not want to over-interpret the statistical significance tests. Equation 2 for the EXAM variable and equations 3 and 4 for the SET variable require only data that can be obtained from the university's records office. All other equations rely on variables that need to be obtained by student questionnaire. Adding student questionnaire variables appears to make little difference for the EXAM equations. However, these variables are rather important for the SET equations. The explanatory power of the SET equations drops significantly if these variables are left out.

There are no surprises in the EXAM equations, except, perhaps, for the fact that HRSWORK is close to being statistically significant and positive, which confirms Lillydahl's (1990) results. The CHOICE variable is highly significant statistically in the SET regressions. HRSWORK enters the SET equation with a negative sign, which confirms the results of Lopus and Maxwell (1995). Also noteworthy for the SET regression is the strongly positive impact of both expected and actual course grade. These results correspond well with those of Marlin and Niss (1980).

Derivation and Interpretation of the Performance Measures

Table 3 reports in lines 1 through 4 the unadjusted EXAM and SET scores for the 11 instructors as well as simple rankings based on these scores. The information provided in lines 3 and 4 of Table 3 would typically form the basis for evaluating the classroom performance of instructors. Instructors B and A are the two top-ranked persons based on the unadjusted SET score. Ordinarily, they would be up for some form of commendation. A look at lines 1 and 2 of Table 3 reveals that the high SET ranking of instructor B is not reflected in a high EXAM ranking. In fact, instructor B's students rank the worst on the objective knowledge test. This is an example of the negative association between student evaluations and student learning that has been reported in the literature (e.g., Abrami et al. 1990, Gramlich and Greenlee 1993).

Given the possibility of a negative correlation between student evaluations and student learning, a weighted score of the two measures could provide an improvement of the assessment of classroom performance. Two alternatively weighted average scores are shown in lines 5 to 8 of Table 3. As the objective EXAM score is weighted more heavily relative to the subjective SET score, it is apparent that instructor B's ranking drops, while the rankings of instructors C and D rise.[12]

The very low EXAM score of instructor B in Table 3 leaves one wondering whether the instructor did not manage to teach his/her students properly or whether there were other factors responsible for the low score that

[11] Compare the probability values reported at the bottom of Table 2. See also the descriptions of the statistical tests in the *Notes* to Table 2.

[12] *Exam scores* are not available for instructors I and K.

TABLE 2. REGRESSION RESULTS FOR THE EQUATIONS PREDICTING EXAM AND SET SCORES

	EXAM Equations		SET Equations			
	Equation 1	Equation 2	Equation 1	Equation 2	Equation 3	Equation 4
C	-8.35	-3.49	3.26	3.21	3.82	2.98
	(-0.77)	(-0.36)	(4.41)	(4.09)	(4.91)	(3.78)
GPA	5.33	5.09	-0.14	-0.15	-0.14	0.10
	(3.78)	(4.18)	(-1.46)	(-1.41)	(-1.33)	(1.06)
ACT	0.92	0.96	-0.03	-0.03	-0.04	-0.01
	(3.77)	(4.43)	(-2.16)	(-2.00)	(-2.15)	(-0.76)
AGE	1.01	1.05	0.01	0.02	0.01	0.04
	(2.93)	(3.22)	(0.32)	(0.71)	(0.41)	(1.56)
MALE	2.28	3.80	0.20	0.23	0.22	0.21
	(1.47)	(2.69)	(1.96)	(2.23)	(2.07)	(1.94)
ATMHRS	0.18	-0.08	0.00	0.00	0.01	-0.01
	(0.56)	(-0.28)	(-0.15)	(0.09)	(0.27)	(-0.26)
REQ	-0.97	2.01	-0.25	-0.18	-0.14	-0.15
	(-0.38)	(1.08)	(-1.40)	(-0.96)	(-0.70)	(-0.75)
NIGHT	-10.30	-9.44	0.42	0.39	0.20	-0.04
	(-2.39)	(-2.32)	(1.27)	(1.13)	(0.58)	(-0.12)
JOB	-2.55		0.34	0.44		
	(-0.87)		(1.78)	(2.19)		
HRSWORK	0.19		-0.01	-0.01		
	(1.89)		(-1.99)	(-1.96)		
CHOICE	4.26		0.49	0.58		
	(1.87)		(3.30)	(3.73)		
EXPECTED			0.44			
			(6.98)			
GRADE				0.30	0.33	
				(4.40)	(4.96)	
R^2	0.2769	0.2712	0.2239	0.1509	0.0992	0.0243
P-values for:						
F-test	0.000	0.000	0.000	0.000	0.000	0.391
Jarque-Bera	0.270	0.336	0.000	0.001	0.000	0.000
Reset	0.910	0.871	0.003	0.259	0.153	0.098
LM-Het	0.749	0.961	0.087	0.007	0.156	0.660

Notes: Parentheses contain t-values. The *F-test* tests for the overall significance of the regression equation. *Jarque-Bera* (Jarque and Bera 1987) tests for the null of normal regression residuals, *Reset* (Ramsey 1969) for the null of no functional form misspecification, and *LM-Het* for the null of homoscedasticity of the regression residuals. The latter is simple LM test where the squared residuals are regressed on a constant term and the squared fitted values.

TABLE 3. PERFORMANCE EVALUATION BASED ON UNADJUSTED EXAM AND SET SCORES, BY INSTRUCTOR

		A	B	C	D	E	F	G	H	I	J	K
1	*EXAM Score*	56.7	50.8	55.9	55.1	51.6	54.4	51.1	52.7		61.4	
2	rank	2	9	3	4	7	5	8	6		1	
3	*SET Score*	4.23	4.29	3.28	3.73	4.13	3.00	3.44	2.67	3.97	4.10	3.90
4	rank	2	1	9	7	3	10	8	11	5	4	6
	Average of EXAM and SET											
5	*Exam weight: α=0.5*	1.092	1.046	0.956	1.010	1.032	0.905	0.934	0.845		1.118	
6	rank by averaged score	2	3	6	5	4	8	7	9		1	
7	*Exam weight: α=0.75*	1.067	0.990	0.992	1.011	0.990	0.952	0.937	0.907		1.123	
8	rank by averaged score	2	6	4	3	5	7	8	9		1	

Notes: Rows 5 and 7 are calculated as: α**Exam Score*/(Mean of *Exam Scores*) + (1-α)**Set Score*/(Mean of *Set Scores*). *Exam scores* are not available for instructors I and K because they did not participate with their classes in the exams.

were outside the influence of the instructor. As suggested earlier, the instructor should be held responsible for a low performance score only to the extent it is not caused by factors outside of his/her influence. Similarly, the instructor should not receive a commendation for good teaching based on a high SET score if the score is the result of grade inflation.

Table 4 provides for each instructor the adjusted scores M_1 and M_2 for both the EXAM and SET scores. The values of M_1 and M_2 are based on the regression results reported in Table 2. The equation numbers identified in Table 4 match those of Table 2. It is apparent that the scores or rankings based on M_1 are not identical to those based on M_2 in all cases. For some instructors, the two measures differ more than for others. This applies in particular to the scores based on the SET equations. For example, for instructor B, there are significant rank differences, while there are none for instructor A. The numbers for instructor B would suggest that numerous students have very negative opinions of this instructor, while instructor A does not elicit any strong reactions in either direction. Instructor I would appear to generate some very positive reactions from students since M_2 is consistently lower than M_1.

Comparing the M_1 and M_2 measures with the unadjusted scores reported in Table 3 one can identify some significant differences. For example, the EXAM rankings of instructors E and G increase markedly in Table 4 relative to Table 3, while instructor D drops in rank. Based on the scores for EXAM equation 1, instructor B's ranking remains at the very bottom. Circumstances beyond the instructor's control can apparently not account for the low test scores of his/her students. Instructor B ranks lower in Table 4 than in Table 3, in particular on the basis of score M_2. It is also apparent that the inclusion of the exam grades in the regression equations (Table 2) makes a difference for instructor B as it lowers his/her adjusted score markedly. This fact suggests some form of grade inflation is likely to be responsible for the good SET scores reported for instructor B in Table 3.

If one wants to use the classroom evaluations to identify good teachers that others can learn from, the results of Table 4 suggest that instructor B should certainly not be among those teachers. A reasonable selection procedure would identify those instructors with (a) high ranks in both the EXAM and the SET category and (b) positive values for M_1 or M_2, whichever is used. In the EXAM category, such a selection would most likely include instructors J and C, and perhaps instructors E and A, depending on the underlying regression equation and the performance measure being used. In the SET category, instructors A and E would likely be the exemplary teachers. Overall, one could end up selecting either instructor A or instructor E, depending on the performance measure that is in use. These two instructors would also likely be selected as exemplary if one used the raw SET scores from Table 3. But in contrast to Table 3, Table 4 clearly eliminates instructor B as an

exemplary instructor. It is important to note that using raw SET scores would not achieve the same result even if university administrators actually do what they routinely say in conversations on the administrative use of raw SET scores, that is, compare an instructor's SET scores over time. Comparing raw SET scores over time will also not help eliminate the persistent upward or downward bias in the performance evaluations of the many instructors that are teaching the same or very similar types of classes most of the time.

TABLE 4. ADJUSTED PERFORMANCE MEASURES AND INSTRUCTOR RANKING FOR EQUATIONS OF TABLE 2

	A	B	C	D	E	F	G	H	I	J	K
EXAM equation 1: M_1	0.083	-0.400	0.129	-0.333	0.208	-0.364	-0.075	-0.286		0.292	
M_2	0.101	-0.620	0.284	-0.437	0.108	-0.356	-0.207	-0.578		0.615	
rank based on M_1	4	9	3	7	2	8	5	6		1	
M_2	4	9	2	7	3	6	5	8		1	
EXAM equation 2: M_1	-0.035	-0.158	0.050	-0.250	0.074	-0.083	-0.058	-0.286		0.345	
M_2	0.099	-0.319	0.267	-0.546	-0.015	-0.336	-0.120	-0.635		1.027	
rank based on M_1	4	7	3	8	2	6	5	9		1	
M_2	3	6	2	8	4	7	5	9		1	
SET equation 1: M_1	0.292	0.200	-0.094	0.167	0.155	-0.545	-0.238	-0.400	0.074	0.083	-0.250
M_2	0.067	0.015	-0.051	0.029	0.029	-0.113	-0.046	-0.182	0.020	0.011	-0.052
rank based on M_1	1	2	7	3	4	11	8	10	6	5	9
M_2	1	5	8	3	2	10	7	11	4	6	9
SET equation 2: M_1	0.271	0.200	-0.219	0.000	0.254	-0.636	-0.095	-0.400	0.074	0.125	-0.125
M_2	0.069	0.011	-0.086	0.025	0.038	-0.108	-0.029	-0.199	0.032	0.009	-0.003
rank based on M_1	1	3	9	6	2	11	7	10	5	4	8
M_2	1	5	9	4	2	10	8	11	3	6	7
SET equation 3: M_1	0.313	0.133	-0.242	0.000	0.239	-0.545	-0.143	-0.450	0.074	0.167	-0.125
M_2	0.053	0.008	-0.100	0.016	0.043	-0.106	-0.026	-0.239	0.037	0.012	-0.001
rank based on M_1	1	4	9	6	2	11	8	10	5	3	7
M_2	1	6	9	4	2	10	8	11	3	5	7
SET equation 4: M_1	0.313	0.400	-0.182	0.167	0.296	-0.364	-0.071	-0.350	0.148	0.167	-0.125
M_2	0.041	0.040	-0.091	0.013	0.043	-0.111	-0.031	-0.259	0.024	0.018	0.000
rank based on M_1	2	1	9	4	3	11	7	10	6	4	8
M_2	2	3	9	6	1	10	8	11	4	5	7

Notes: Instructors are listed by column. Instructors I and K did not participate in the common final exam.

If one uses a formula approach, such as based on the measure *P*, the results could look like the ones in Table 5. In this table, a number of alternative performance measures are provided to check the sensitivity of the performance measure with regard to (a) the information requirements, as represented by the variables in the underlying regression equations, (b) the weight that is given to the EXAM score relative to the SET score, and (c) the choice of measure M_1 versus M_2. Regardless of any of these measurement choices, instructor J turns out to receive the best performance evaluation. His/her adjusted EXAM scores are so much better than those of the

other instructors that they easily compensate for his/her lower SET scores. This is what can happen with a formalized approach that simply averages scores. One notices that instructors A and E are also likely candidates that could be selected as exemplary teachers based on most of the results of Table 5. Hence, the formal averaging procedure produces results that are similar to the more informal approach that relies on finding candidates with both high SET and EXAM scores.

TABLE 5. OVERALL PERFORMANCE EVALUATION BASED ON ADJUSTED EXAM AND SET SCORES (P_J)

	A	B	C	D	E	F	G	H	I	J
EXAM equation 1 and SET equation 1, weight α = 0.5										
based on M_1	0.188	-0.100	0.018	-0.083	0.182	-0.455	-0.157	-0.343		0.188
rank	2	6	4	5	3	9	7	8		1
based on M_2	0.084	-0.302	0.116	-0.204	0.068	-0.235	-0.126	-0.380		0.313
rank	3	8	2	6	4	7	5	9		1
EXAM equation 2 and SET equation 3, weight α = 0.5										
based on M_1	0.139	-0.012	-0.096	-0.125	0.157	-0.314	-0.100	-0.368		0.256
rank	3	4	5	7	2	8	6	9		1
based on M_2	0.076	-0.155	0.083	-0.265	0.014	-0.221	-0.073	-0.437		0.519
rank	3	6	2	8	4	7	5	9		1
EXAM equation 1 and SET equation 1, weight α = 0.75										
based on M_1	0.135	-0.250	0.073	-0.208	0.195	-0.409	-0.116	-0.314		0.240
rank	3	7	4	6	2	9	5	8		1
based on M_2	0.093	-0.461	0.200	-0.320	0.088	-0.295	-0.167	-0.479		0.464
rank	3	8	2	7	4	6	5	9		1
EXAM equation 2 and SET equation 3, weight α = 0.75										
based on M_1	0.052	-0.085	-0.023	-0.188	0.115	-0.199	-0.079	-0.327		0.300
rank	3	6	4	7	2	8	5	9		1
based on M_2	0.087	-0.237	0.175	-0.405	0.000	-0.278	-0.096	-0.536		0.773
rank	3	6	2	8	4	7	5	9		1

Notes: Instructors I and K did not participate in the common final exam. Hence, no weighted score is available.

A Note on Practical Implementation

As for the practical implementation of the assessment methodology in a university environment, it seems that one could divide the path toward more useful teaching evaluations into two steps. In a first step, the suggested regression-based methodology of identifying outstanding instructors could be implemented for teaching evaluations alone. This requires very few additional resources. However, it does require a look at the issue of protecting student confidentiality. Since students need to provide their name and, perhaps, some personal data (for some regressions of Table 2), some mechanism needs to be put in place to safeguard this information. Once this issue is resolved, the suggested methodology can be applied across the board to all classes across the university regardless of field and level. Assessment could be done on the basis of a single regression equation for the whole institution. The only potential modification over the current paper could be the inclusion of dummy variables to account for potential differences in graduate/undergraduate classes, by field of study, or by

college within the university. Compared to today's practice of looking at raw unadjusted student evaluations, adjusted performance scores would be more comparable and, hence, more useful across instructors, classes, and fields. Last but not least, it would make teaching evaluations more respectable among faculty members because the adjusted SET measures would be less vulnerable to the charge that teachers "buy" good teaching evaluations with good grades.

In a second step, one could develop knowledge-based outcomes tests for each field of study to construct measures of performance evaluation similar to the EXAM scores in this paper. This would take more time to implement and, for practical purposes, this step may be limited to principles classes or other large undergraduate classes. In developing cognitive outcomes tests, it would be important to make sure that the instructors do not know the test content so they are not teaching to the outcomes test. This would imply that the questions on the test change each term and that they are made up by an instructor not teaching a class during the term. Also, all instructors need to allocate the same percentage of the final course grade to this outcomes test to provide all students with the same incentive. Once all instructors with significant teaching duties teach at least one of the courses evaluated with an outcomes test, an instructor's overall teaching effectiveness could be evaluated by the methodology suggested in this study.

It should be apparent that the suggested methodology can be implemented with a few adaptations also in a non-university environment where employees subjectively evaluate their managers on a regular basis and managers are also evaluated by some objective outcomes measure, such as profit. A regression-based approach is clearly better suited for larger corporations than for very small companies. But then again, larger companies are also more likely to have a formalized performance evaluation in place. As in the case discussed in this study, a good understanding of the environment in which the evaluation is taking place is needed to identify the appropriate variables for the regression equations. In addition, safeguarding employees' evaluations of their superiors would require some careful attention.

Summary and Conclusions

The purpose of this paper has been to suggest a formal methodology to adjust regular performance evaluations for a number of factors that can bias them. One key set of factors includes the circumstances that are beyond the control of those being evaluated. Removing bias from performance evaluations is important because persistent bias in the evaluation process does not tend to remain unnoticed among those being evaluated. It can contribute to making the evaluation process a matter of disdain or ridicule and ultimately lead to a climate where striving for excellence is considered not worth it any more.

The methodology has been demonstrated for the case of evaluating university instructors' in-class teaching performance. The example has revealed that the methodology can identify exemplary teachers and distinguish them from those instructors that obtain high student evaluations by grade inflation. The example has also revealed that traditional arguments in defense of using simple student evaluations for measuring teaching performance are unfounded. Specifically, looking at teaching evaluations over time cannot typically eliminate the bias identified and targeted in this study.

The suggested methodology does require some changes compared to the currently popular system of student evaluations. The major change is that more data need to be collected. In particular, students need to reveal their identity so that student specific data can be incorporated in the evaluation process. Although that would be unfamiliar territory, it should not pose an insurmountable problem if the purpose is explained thoroughly to all involved and administrative safeguards are in place against the abuse of this information.

As emphasized at the outset of this paper, the suggested methodology of performance evaluation is not limited in its potential usefulness to evaluating college teaching. On the contrary, the basic idea of making those being evaluated responsible only for those outcomes that they can control is applicable in most environments where formal performance evaluations are conducted. One would expect that performance evaluations that are based on this idea would be considered more acceptable by those being evaluated. Achieving a higher acceptance rate would be no small change. It would make it considerably easier to communicate or understand why changes in behavior are needed. That way, performance evaluations could get a step closer to their ultimate objective: fine-tuning behavior to raise efficiency.

References

Abrami, P.C., S. d'Apollonia, and P.A. Cohen. 1990. "Validity of Student Ratings of Instruction: What We Know and What We Do Not." *Journal of Educational Psychology* 82: 219-231.

Bailey, Charles D., Sanjay Gupta, and Richard W. Schrader. 2000. "Do Students' Judgment Models of Instructor Effectiveness Differ by Course Level, Course Content, or Individual Instructor?" *Journal of Accounting Education* 18: 15-34.

Boex, L.F. Jameson. 2000. "Attributes of Effective Economics Instructors: An Analysis of Student Evaluations," *Journal of Economic Education* 31: 211-227.

Brasfield, David W., Dannie E. Harrison, and James P. McCoy. 1993. "The Impact of High School Economics on the College Principles of Economics Course." *Journal of Economic Education* 24: 99-111.

Cascio, Wayne F. 1998. *Applied Psychology in Human Resource Management* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Dilts, David A. 1980. "A Statistical Interpretation of Student Evaluation Feedback." *Journal of Economic Education* 11: 10-15.

_____, and Ali M. Fatemi. 1982. "Student Evaluation of Instructors: Investment or Moral Hazard?" *Journal of Financial Education* 12: 67-70.

Gramlich, Edward M., and Glen A. Greenlee. 1993. "Measuring Teacher Performance." *Journal of Economic Education* 24: 3-13.

Hansen, W. Lee, and Allen C. Kelley. 1973. "Political Economy of Course Evaluations." *Journal of Economic Education* 4: 10-21.

Jarque, Carlos M., and Anil K. Bera. 1987. "A Test for Normality of Observations and Regression Residuals." *International Statistical Review* 55: 163-72.

Kreitner, Angelo, and Angelo Kinicki. 2000. *Organizational Behavior and Management*. New York: McGraw-Hill.

Leventhal, Les, Philip C. Abrami, Raymond P. Perry, and Lawrence J. Breen. 1975. "Section Selection in Multi-Section Courses: Implications for the Validation and Use of Student Rating Forms." *Educational and Psychological Measurement* 35: 885-895.

Lopus, Jane S., and Nan L. Maxwell. 1994. "Beyond High School: Does the High School Economics Curriculum Make a Difference?" *The American Economist* 38: 62-69.

_____. 1995. "Teaching Tools: Should We Teach Microeconomic Principles Before Macroeconomic Principles?" *Economic Inquiry* 33: 336-350.

Lillydahl, Jane H. 1990. "Academic Achievement and Part-Time Employment of High School Students." *Journal of Economic Education* 21: 307 - 316.

Marco, G.L., A.A. Abdel-Fattah, and P.A. Baron. 1993. "Methods Used to Establish Score Compatibility on the Enhanced ACT Assessment and the SAT." College Board Report No. 92-3, New York.

Marlin Jr., James W., and James F. Niss. 1980. "End-of-Course Evaluations as Indicators of Student Learning

and Instructor Effectiveness." *Journal of Economic Education* 11: 16-27.

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

Mirus, Rolf. 1973. "Some Implications of Student Evaluations of Teachers." *Journal of Economic Education* 4: 35-37.

Nelson, Jon P., and Kathleen Lynch. 1984. "Grade Inflation, Real Income, Simultaneity, and Teaching Evaluations." *Journal of Economic Education* 15: 21-37.

Ramsey, James B. 1969. "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis." *Journal of the Royal Statistical Society, ser. B*, 31: 350-71.

Rose, Louis A. 1975. "Adjustment of Student Ratings of Teachers for Extrinsic Influences." *Journal of Economic Education* 6: 129-132.

Seiver, Daniel A. 1982. "Evaluations and Grades: A Simultaneous Framework." *Journal of Economic Education* 13: 32-38.

Shmanske, Stephen. 1988. "On the Measurement of Teacher Effectiveness." *Journal of Economic Education* 19: 307-314.

Simpson, Penny M., and Judy A. Siguaw. 2000. "Student Evaluations of Teaching: An Exploratory Study of the Faculty Response." *Journal of Marketing Education* 22: 199-213.

Tuckman, Howard P. 1975. "Teacher Effectiveness and Student Performance." *Journal of Economic Education* 6: 34-39.

Watts, Michael and Gerald J. Lynch. 1989. "The Principles Courses Revisited." *American Economic Review* 79: 236-241.

Weidenaar, Dennis, J., and Joe A. Dodson, Jr. 1972. "The Effectiveness of Economics Instruction in Two-Year Colleges." *Journal of Economic Education* 3: 5-12.

Zangenehzadeh, Hamid. 1988. "Grade Inflation: A Way Out." *Journal of Economic Education* 19: 217-226.