1975

Leniency, Learning, and Evaluations

John P. Palmer

Geoffrey Carliner

Thomas Romer

20236

RESEARCH REPORT 7516

by

John Palmer
University of Western Ontario

Geoffrey Carliner
University of Western Ontario
and
Thomas Romer
Carnegie-Mellon University

July 1975

LENIENCY, LEARNING, AND EVALUATIONS

by

John Palmer
University of Western Ontario

Geoffrey Carliner
University of Western Ontario

and

Thomas Romer
Carnegie-Mellon University

July 1975

The authors thank the members of the University of Western Ontario, Department of Economics for their cooperation and assistance with this project. We are also grateful for helpful suggestions from Dennis Capozza, Kenneth Doyle, and Donald Hoyt.

Leniency, Learning, and Evaluations

1. Introduction

With student evaluations of instructor effectiveness playing an increasingly important role in the determination of merit pay, promotion, and tenure, there is a growing interest in what these evaluations actually measure. Faculty members frequently voice doubts about using student evaluations because it is not clear to what extent they measure the leniency of the instructors, the amount the instructors taught the students, or the performing ability of the instructors.

1.1 Evaluations and Leniency

Several recent studies have documented a positive relationship between the grades economics students receive and the evaluations they give their instructors (Kelley, 1972; Capozza, 1973). Similar results have also been reported for other disciplines (Murray, 1972) and across various disciplines (Nichols and Soper, 1972; Perry and Bauman, 1973; Reuber, 1974). These results are consistent with the view that instructors "buy" high evaluations (and, they hope, higher pay, promotion, and tenure) by "giving" the students higher grades. This view suggests that there is at least a tacit collusion between instructors and students to scratch each other's backs. The results are also consistent, though, with several other behavioural models. Students with higher grades may have given higher evaluations to their instructors because the instructors in these samples taught

to the brighter students. Alternatively, it is possible that a positive correlation between grades and evaluations could be observed if the better instructors, who justifiably received higher evaluations, taught their students more, so that their students justifiably earned higher grades. Finally, the causation may be in the opposite direction from that usually assumed, and "...an instructor might grade a class harshly or generously because of the ratings he receives (or anticipates)." (Doyle, 1974.)

Many other studies have found no relationship between grades and evaluations. These studies are well summarized by Costin et al. (1971) and Menges (1973). But as McKenzie and Tullock (1975) point out, the lack of a correlation between grades and evaluations does not necessarily lead to a rejection of the hypothesis that more lenient instructors receive higher evaluations. If instructors become more lenient in the hopes of receiving higher evaluations, the students may respond simply by studying less, and learning less, yet receiving no lower grades. This phenomenon is particularly likely if students value additional leisure time highly and are satisficers with respect to grades. As a result, the use of grades, uncorrected for the knowledge obtained by the students, as a measure of instructor leniency may be quite misleading.

1.2 Evaluations and Amount Learned

Attempts to measure the relationship between learning and student evaluations of instructor effectiveness have yielded mixed results. Capozza (1973) reported a negative and significant relationship between evaluations and the amount students learned, but he has since then indicated to us by correspondence that with a larger sample his results are no longer statistically significant. Besides using grades as a measure of leniency, which we have already suggested may be inappropriate, Capozza also failed to include

any variables in his model to explain why some students might learn more than others. Rodin and Rodin (1972) also found a significantly negative relationship between evaluations and the amount learned, but their study has been found lacking in several respects (see Frey, 1973 and Eble, 1974) including small sample size and omitted variables.

Crowley and Wilton (1974) found a positive but insignificant relationship between some components of evaluations and the amount students learned in beginning economics. Significantly positive relationships have been reported by Gessner (1973), Frey (1973), and Doyle and Whitely (1974).

It appears from the studies which have previously been conducted and from the criticisms leveled at them that the issues have been clouded by rhetoric and by the complexity of the relationships. What is needed is a model which measures, first, the impact of the instructor on the amount his students learn, correcting for other possible influences on learning. Second, the model must measure the leniency of the instructor, correcting for other influences (including the amount learned) on students' grades. And third, it must relate these measures to students' evaluations of instructor effectiveness, correcting for other possible influences. What is needed, then, is a sequential, three-equation model to determine the effects of learning and leniency on evaluations. We turn now to our development of such a model.

2. Specification

2.1 Important Variables

The knowledge of economic concepts (KNOW) gained by a student in the microeconomic portion of a beginning economics course depends on many things, most of which are quantifiable. A list of these factors includes:

(1) Previous knowledge of economics concepts (PRE). Students knowing more economics at the beginning of a course may well know more than others at the end of the course, though they may not learn as much new material during the course.

(2) Amount of previous economics (PE). If the student has had an economics course previous to this one (perhaps in high school or perhaps a university course which he or she failed), we would expect the student to know more at the conclusion of the course.

(3) Amount of calculus taken by the student (CALC). Because much of microeconomic theory explicitly or implicitly deals with differentiation and integration, students with a calculus background may learn the concepts more easily than students without a calculus background. The amount of calculus in a student's background may also be a proxy for analytical and mathematical aptitude. (The latter was found by Crowley and Wilton (1974) to have a significantly positive effect on the amount of economics learned by students in beginning courses.)

(4) Previous academic average (AA). Students who have done well in the past in terms of their grades tend to continue to do well, either because of high aptitude or because of high motivation. Ability to take tests is a skill in itself; high academic average is, in part, a reflection of this ability. Because academic averages in secondary schools are probably not commensurate with academic averages for upperclass students at university, we have split AA into two parts: AAF to represent previous academic average of first-year students and AAU the previous academic average of upperclass students.

(5) Academic year of the student (Y). If upperclass students are more mature than first-year students they may learn more in a course. It is also possible, however, as suggested by Crowley and Wilton, that upperclass students view a beginning economics course as one which deserves less of their attention and effort, so that they learn less. Also, students who postpone taking their first economics course until their 2nd or 3rd year in university may have less aptitude for it than first-year students.

(6) Time the class meets (T). Students might learn more in classes meeting at certain times of day than they would from classes meeting at other times of the day.

(7) Size of the class (SZ). We include this variable to see if class size actually affects learning.

(8) Sex of the student (FEM). Crowley and Wilton found that female students learn significantly less in a beginning economics course than male students do. Their measure of amount learned, however, was biased against students beginning the course with less knowledge of economic concepts, so that if females began a course knowing less economics and improved their knowledge by the same absolute amount as males did, then the Crowley and Wilton measure of amount learned would yield their result spuriously. We are including a dummy variable for females to determine whether the knowledge a student has of economic concepts in terms of absolute raw scores varies with the sex of the student, ceteris paribus.

Students in fourteen sections of the microeconomics portion of the Principles of Economics course at the University of Western Ontario were given a 19-question multiple-choice examination at the beginning of their first class in September, 1974.[1] This examination is the "pretest." The examination was administered by persons not teaching the course and instructors of the course were not permitted to see the questions on the examination. Examination questions were designed to test students' mastery of economic concepts rather than of economic jargon. This pretest serves as our measure of PRE, a student's previous knowledge of economics. The same examination was given to these students under examination conditions at the end of the term in December. This "post-test" is used as our measure of KNOW, a student's current knowledge of economics.[2]

---

[1] Principles of Economics is taught at U.W.O. in many sections, with an average enrolment of about 60 students per section.

The examination we used was a slightly modified version of the microeconomics portion of a test, similar in nature to the American TUCE but better suited for testing Canadian students, designed by Crowley and Wilton (1974). We eliminated some questions found to be fairly weak indicators of student knowledge by Crowley and Wilton and added a few questions to cover omitted material we felt ought to be included. A copy of the exam is available from the authors.

[2] Students who dropped the course were omitted from the sample, as were those who changed sections. Our final sample included 617 students.

After the "post-test" was administered, we asked the instructors to indicate the degree of correspondence between the material covered by the "post-test" and material covered in class. This correspondence was found to be uniformly high for all sections, so that we are fairly confident that our test measures areas of knowledge covered in all sections in the sample.[3]

Students in sections in which multiple-choice testing is used regularly throughout the term may not know more economics than others in their cohort but may simply have had better training in taking economics multiple-choice exams and may therefore do better on our tests. It seems appropriate to control for this possibility by including an additional variable, viz.:

(9) Previous experience with multiple choice questions in economics (MULT).

2.2 "Knowledge" equation

In order to gauge an instructor's contribution to student knowledge of economics, we estimated the following equation:

Equation 1 i = student, j = section

$$KNOW_{ij} = \alpha_o + \alpha_1 PRE_{ij} + \alpha_2 PE_{ij} + \alpha_3 CALC_{ij} + \alpha_4 AAF_{ij} + \alpha_5 AAU_{ij} + \alpha_6 Y_{ij} + \alpha_7 T_j + \alpha_8 SZ_j + \alpha_9 FEM_{ij} + \alpha_{10} MULT_{ij} + \alpha_{11} INST_j + u_{KNOW}$$

---

[3] An instructor whose class material differed significantly from that covered on the "post-test" may have taught his students as much economics as did other instructors but his students would not have done as well, ceteris paribus, on the post-test. The uniformly high degree of correspondence between post-test and material covered in class is therefore reassuring. In large measure, this result is probably due to the use of a common text and reading list in this course.

With the exception of INST, the variables in this equation have been defined above. In the estimation, we have treated the variables as dummies--the precise definition of these dummy variables is given in Section 3.

INST is a set of dummy variables, one each for all but one instructor who serves as a kind of "numéraire". The set of estimated coefficients $\hat{\alpha}_{11}$ thus gives us an estimate of the contribution of each instructor to students' knowledge, relative to the contribution of the omitted teacher. A high value of $\hat{\alpha}_{11}$ will be associated with an instructor whose contribution to student knowledge is relatively great, while an instructor with a relatively small contribution will have a low $\hat{\alpha}_{11}$.

2.3 "Leniency" equation

In order to determine the extent of an instructor's leniency in assigning grades to students, we must control for variables other than leniency which may affect each student's grade. Aside from instructor leniency, the grade a student receives[4] (GRADE) will depend on the variables (2) through (9), defined in Section 2.1, as well as on the amount that the student knows, which we measure by KNOW. Consequently, we estimated the following equation:

Equation 2

$$\begin{aligned}\text{GRADE}_{ij} = \beta_o + \beta_1 \text{PE}_{ij} + \beta_2 \text{CALC}_{ij} + \beta_3 \text{AAF}_{ij} + \beta_4 \text{AAU}_{ij} + \beta_5 \text{Y}_{ij} + \beta_6 \text{T}_j \\ + \beta_7 \text{SZ}_j + \beta_8 \text{FEM}_{ij} + \beta_9 \text{KNOW}_{ij} + \beta_{10} \text{MULT} + \beta_{11} \text{INST}_j + u_{\text{GRADE}}\end{aligned}$$

In this equation, the set of estimated coefficients $\hat{\beta}_{11}$ play a role analogous to that of $\hat{\alpha}_{11}$ in Equation (1). Here the coefficients of INST provide a measure of the relative leniency of each instructor, net of the leniency of the numéraire teacher. High values of $\hat{\beta}_{11}$ will be associated with

[4] Grades are assigned on a numerical scale with 100 as the maximum.

relatively more lenient instructors.[5]

Because instructors and other variables are expected to have an impact on students' knowledge, including these same variables along with KNOW (measured by the post-test scores) in the regressions may create problems of multicollinearity and bias the estimated coefficients. An alternative specification of Equation 2 is to substitute for KNOW from Equation 1:

Equation 2A

$$\begin{aligned} GRADE_{ij} = {} & (\beta_o + \beta_9 \alpha_o) + (\beta_1 + \beta_9 \alpha_2) PE_{ij} \\ & + (\beta_2 + \beta_9 \alpha_3) CALC_{ij} + (\beta_3 + \beta_9 \alpha_4) AAF_{ij} \\ & + (\beta_4 + \beta_9 \alpha_5) AAU_{ij} + (\beta_5 + \beta_9 \alpha_6) Y_{ij} \\ & + (\beta_6 + \beta_9 \alpha_7) T_j + (\beta_7 + \beta_9 \alpha_8) SZ_j \\ & + (\beta_8 + \beta_9 \alpha_9) FEM_{ij} + \beta_9 \alpha_1 PRE_{ij} \\ & + (\beta_{10} + \beta_9 \alpha_{10}) MULT_{ij} + (\beta_{11} + \beta_9 \alpha_{11}) INST_j \\ & + (\mu_{GRADE} + \beta_9 \mu_{KNOW}) . \end{aligned}$$

β_9 can be estimated by dividing the coefficient of PRE by α_1 from Equation 1. With this estimate of β_9 and the estimates of the α_i's, the remaining $\hat{\beta}$'s can be disentangled.

In addition to providing us with information concerning instructor leniency and contribution to student knowledge, Equations (1) and (2) can be

---

[5] What we are really interested in, of course, is the students' perception of instructor leniency. Because perceived leniency may not be closely related to final grades in the course, in estimating Equations (2) and (2A) we used each student's grade in the course just prior to the time the evaluations were conducted. The teaching evaluations were carried out approximately two-and-a-half weeks prior to the end of the term's lectures.

used to see whether variables such as sex of student, calculus, time of class, student's year, etc. have effects on student knowledge of economics different from--perhaps even opposite to--their impact on the student's grade in the course.

2.4 "Evaluation" equation

We can use the estimated coefficients $\hat{\alpha}_{11}$ and $\hat{\beta}_{11}$ from Equations (1) and (2) to explore the relative importance of the instructor's teaching ability and the average leniency of an instructor in determining the student evaluation of that instructor. The evaluation questionnaire included an "overall effectiveness" question: "How would you rate your instructor in terms of general, overall effectiveness as a teacher?" Students were asked to give their ratings on an integer scale ranging from 5 ("Outstanding") to 1 ("Poor").[6]

It would be most desirable, for the purposes of our experiment, to identify each student's evaluation of his instructor with the students own knowledge and grade. Unfortunately, this was not possible, because the evaluations were done anonymously.[7] As a result, we were forced to use section averages for our regressions involving student evaluations of the instructors. These section averages are denoted by E_j. Our third equation is:

Equation 3

$$E_j = \gamma_o + \gamma_1 \hat{\alpha}_{11_j} + \gamma_2 \hat{\beta}_{11_j} + u_E \qquad j=1,\ldots,14$$

The independent variables in this equation are the estimated coefficients on contribution to learning (from Equation (1)) and instructor leniency (from

---

[6]The gradations are:
5- Outstanding
4- Very good
3- Good
2- Satisfactory
1- Poor

[7]In the past, students at U.W.O., fearing reprisals from their instructors, refused to identify themselves with their student numbers on evaluation forms. This resulted in a high incidence of invalid responses, and the solicitation of student number was abandoned in 1974.

Equation (2) or (2A)). It should be noted that in this study we are not attempting to explain all the factors that go into the determination of student evaluations of instructors. Our aim is more modest. The estimates of Equation (3) will indicate whether or not the amount taught to students by an instructor and the instructor's leniency in "handing out" grades have a statistically significant influence on student ratings of instructors and, if so, which effect is stronger.

3. The Results

The model described in the previous section was estimated using ordinary least squares. In this section, we discuss these results, focusing first on the estimates of Equations 1, 2, and 2A and then on Equation 3.

3.1 Knowledge and Reward

Our estimates of Equations 1, 2, and 2A are presented as Regressions 1, 2, and 2A, respectively. In these regressions, all the independent variables are entered as dummy variables, whose definitions are given in Table 1. We believe that the results provide some interesting information about the factors influencing a student's knowledge of economics at the end of a semester of micro principles and the grade a student receives. Since the regressions have most of their explanatory variables in common, it seems natural to discuss the results in terms of the impact of each of these variables.

Previous economics. It appears that having had an economics course prior to the college principles course has at best no effect on a student's knowledge or his grade in the principles course. Having had previous economics may even have an adverse effect on both KNOW and GRADE. In Regression 1, the coefficients on PE2, PE3, and PE4 (student had some previous economics) are all negative but insignificant, while in Regression 2 the coefficient of PE1 (student had no previous economics) is positive and significant at the 10% level. Since nearly all of those students who say they have had "economics" prior to the principles course had such a course in secondary school, these results may shed some light

Table 1

Definitions of Variables

Variable	value of variable = 1, if...
PE1	no previous economics course
PE2	one previous economics course, passed
PE3	one previous economics course, failed
PE4	more than one previous economics course
CALC1	no previous calculus course
CALC2	one term of previous calculus
CALC3	two terms of previous calculus
CALC4	more than two terms of previous calculus
AAFA(AAUA)	previous academic average of A, freshman (upperclassman)
AAFB(AAUB)	previous academic average of B, freshman (upperclassman)
AAFC(AAUC)	previous academic average of C, freshman (upperclassman)
AAFD(AAUD)	previous academic average of D, freshman (upperclassman)
Y1	first-year student
Y2	second-year student
Y3	third-year student and other
FEM	1 = female student, 0 = male
MULT1	classroom tests and assignments < 25% multiple choice
MULT2	classroom tests and assignments 26-50% multiple choice
PRE2	2 or fewer correct answers on pretest
PRE3	3 correct answers on pretest
PRE4	4 correct answers on pretest
⋮	⋮
PREK	K correct answers on pretest (3 < K < 12)
⋮	⋮
PRE12	12 or more correct answers on pretest
KDUM6	6 or fewer correct answers on post-test
KDUM7	7 correct answers on post-test
⋮	⋮
KDUMN	N correct answers on post-test (7 < N < 16)
⋮	⋮
KDUMN16	16 or more correct answers on post-test
GRADE	student's course grade just prior to the evaluation
KNOW	score on post-test, 0 - 19

Regression 1 (standard errors in parentheses)

$$\begin{aligned}\text{KNOW} = 7.04 &\underset{(.889)}{} - \underset{(.265)}{0.228}\ \text{PE2} - \underset{(2.78)}{0.429}\ \text{PE3} - \underset{(.518)}{0.437}\ \text{PE4} - \underset{(.535)}{1.23}\ \text{CALC2} + \underset{(.244)}{0.207}\ \text{CALC3} \\ &+ \underset{(.377)}{0.377}\ \text{CALC4} + \underset{(.351)}{2.27}\ \text{AAFA} + \underset{(.266)}{0.804}\ \text{AAFB} + \underset{(1.00)}{1.92}\ \text{AAFD} + \underset{(.740)}{1.06}\ \text{AAUA} \\ &+ \underset{(.475)}{1.65}\ \text{AAUB} - \underset{(.478)}{.018}\ \text{AAUC} + \underset{(.888)}{.620}\ \text{AAUD} + \underset{(.581)}{.070}\ \text{Y3} - \underset{(.253)}{0.176}\ \text{FEM} \\ &- \underset{(.372)}{0.166}\ \text{MULT1} - \underset{(.355)}{0.064}\ \text{MULT2} + \sum_{k=3}^{12} a_k \text{PREk} + \sum_{k=2}^{14} b_k \text{INST}_k\end{aligned}$$

$$R^2 = 0.311$$

$a_3 = 0.665$ (.788)

$a_4 = 0.661$ (.725)

$a_5 = 1.33$ (.707)

$a_6 = 1.37$ (.690)

$a_7 = 2.02$ (.716)

$a_8 = 2.06$ (.699)

$a_9 = 3.06$ (.734)

$a_{10} = 3.36$ (.776)

$a_{11} = 4.28$ (.971)

$a_{12} = 5.42$ (.875)

$b_1 = 0$ (omitted instructor)

$b_2 = 1.52$ (.524)

$b_3 = 0.609$ (.587)

$b_4 = 1.02$ (.540)

$b_5 = 2.17$ (.607)

$b_6 = 1.42$ (.615)

$b_7 = 2.23$ (.699)

$b_8 = 1.30$ (.566)

$b_9 = 0.632$ (.582)

$b_{10} = 3.00$ (.517)

$b_{11} = 0.258$ (.542)

$b_{12} = 1.34$ (.542)

$b_{13} = 1.84$ (.663)

$b_{14} = 1.83$ (.596)

Regression 2 (standard errors in parentheses)

$$\begin{aligned}\text{GRADE} = {} & \underset{(2.89)}{52.00} + \underset{(1.08)}{1.78}\,\text{PE1} - \underset{(.968)}{4.20}\,\text{CALC1} + \underset{(1.53)}{10.76}\,\text{AAFA} + \underset{(1.13)}{3.22}\,\text{AAFB} - \underset{(3.96)}{4.56}\,\text{AAFD} \\ & + \underset{(3.10)}{14.06}\,\text{AAUA} + \underset{(2.00)}{6.56}\,\text{AAUB} - \underset{(2.00)}{3.69}\,\text{AAUC} + \underset{(3.74)}{1.42}\,\text{AAUD} + \underset{(2.45)}{2.31}\,\text{Y3} \\ & + \underset{(1.06)}{1.26}\,\text{FEM} - \underset{(1.56)}{1.68}\,\text{MULT1} - \underset{(1.50)}{1.38}\,\text{MULT2} + \sum_{n=7}^{16} c_n\,\text{KDUM}_n + \sum_{k=2}^{14} d_k\,\text{INST}_k\end{aligned}$$

$$R^2 = 0.404$$

$c_7 = -0.208$ (2.52)

$c_8 = 4.93$ (2.30)

$c_9 = 4.69$ (2.23)

$c_{10} = 5.36$ (2.16)

$c_{11} = 9.46$ (2.29)

$c_{12} = 10.29$ (2.22)

$c_{13} = 11.34$ (2.22)

$c_{14} = 12.22$ (2.60)

$c_{15} = 16.80$ (2.50)

$c_{16} = 20.05$ (2.68)

$d_1 = 0$ (omitted instructor)

$d_2 = 3.14$ (2.22)

$d_3 = 3.61$ (2.48)

$d_4 = 5.57$ (2.28)

$d_5 = 3.64$ (2.59)

$d_6 = 2.81$ (2.60)

$d_7 = 0.944$ (2.24)

$d_8 = 3.41$ (2.37)

$d_9 = 9.36$ (2.44)

$d_{10} = 0.039$ (2.24)

$d_{11} = 10.97$ (2.27)

$d_{12} = 5.01$ (2.30)

$d_{13} = 5.47$ (2.82)

$d_{14} = 0.492$ (2.54)

Regression 2A (disentangled coefficients)

$$\begin{aligned}\text{GRADE} = 35.18 &+ 1.84\ \text{PE1} - 4.06\ \text{CALC1} \\ &+ 8.90\ \text{AAFA} + 2.48\ \text{AAFB} - 5.04\ \text{AAFD} \\ &+ 12.11\ \text{AAUA} + 4.91\ \text{AAUB} - 3.86\ \text{AAUC} \\ &+ 0.34\ \text{AAUD} + 2.21\ \text{Y3} + 1.58\ \text{FEM} \\ &- 1.53\ \text{MULT1} - 1.50\ \text{MULT2} + 2.43\ (\text{estimated knowledge}) \\ &+ \sum_{k=2}^{14} f_k\ \text{INST}_k\end{aligned}$$

$R^2 = .303$ for the estimated equation

$f_1 = 0.00$

$f_2 = 1.98$

$f_3 = 3.53$

$f_4 = 5.60$

$f_5 = 2.71$

$f_6 = 1.10$

$f_7 = -0.54$

$f_8 = 2.59$

$f_9 = 8.56$

$f_{10} = -1.85$

$f_{11} = 10.70$

$f_{12} = 4.22$

$f_{13} = 4.20$

$f_{14} = -0.56$

on the teaching and learning of secondary-school economics. A student may take a high school course that is billed as an economics course but which, in fact, bears only a vague resemblance to the course he encounters in college. The resemblance is not strong enough to help the student perform better in the college course and may even result in confusing him. A related possibility is that the student is taught a principles course in high school and is taught badly. Alternatively, a student may arrive in the college course with some knowledge but a false sense of having already mastered the material. In either case, his performance in the college course would be adversely affected.

Academic average. Students with academic averages of A or B (prior to enrolling in the principles course) do significantly better both on our post-test and in the principles course than those with lower averages.[9] Upperclass A and B students appear to get higher grades than freshmen in their section with similar knowledge and academic background. This is probably due to the higher standards in university (an A average in college generally represents somewhat better performance than it does in secondary school) and to the greater experience upperclass students have in taking college-level exams. Somewhat surprising is the insignificant coefficient of AAUA in Regression 1--upperclass students with an A average do not know significantly more economics at term's end than do freshmen with a high-school C. Yet the more senior A student can expect a considerably higher grade in the course than a first-year student in his section with a C average and the

---

[8]Something of a puzzle is the positive and significant (at the 5% level) coefficient of AAFD in Regression 1. We have no entirely convincing explanation why freshmen coming in with a D average should do 2 points better, ceteris paribus, on the post-test than those in their cohort with a C average. Perhaps, being underdogs, they try harder. In any case, those in the AAFD category represent a very small fraction (1.3%) of our sample. This result may therefore be due to extraordinary performance by two or three students.

same knowledge! Ability in writing college-level exams appears to be handsomely rewarded.

Sex of student. An interesting non-result is the fact that male and female students of like background do not differ significantly in their performance either on the post-test or in the course itself. Controlling for pre-test performance and academic background, as we did in Regression 1, gives a negative but quite insignificant coefficient on FEM. Similar control in the GRADE equation produces a small, positive, and again quite insignificant coefficient on FEM.[9]

Calculus background. Students who have had no calculus course do slightly (but statistically significantly) better on our post-test than do students who have had a term of calculus, ceteris paribus. Those with even more calculus background do not know significantly more economics at the end of the micro term of principles than do students without any calculus. On the other hand, the lack of a calculus background does work to a student's detriment when it comes to performance in the principles course (cf.: negative coefficient of CALC1 in Regression 2).

Our post-test attempts to measure primarily knowledge of and ability to deal with basic economic concepts and does not reward analytical ability per se. Lectures and course tests, on the other hand, may be more directly concerned with the manipulation of tools of analysis and hence reward more

---

[9]It might be noted that all but one of the instructors in our sample are male, while 28.9% of the students are female.

highly those who have greater exposure to calculus--even though calculus was not explicitly required in handling the problems. Although those without calculus background appear to have at least as good--possibly even better--knowledge of economic concepts as their more numerate classmates, they are at a disadvantage in the course exams and assignments.

Time and size of class. These variables were dropped from the regression by our regression package. (None of the coefficients associated with any of the time and size variables was significantly different from zero at the 99.999% level.) Neither student knowledge nor grade are affected by the time of day that a class meets or whether the class is held in one- or two-hour meetings.

Pre-test and post-test. Students who enter the principles course knowing some economics do significantly better on the post-test than those who know very little at the start. (This can be seen in Regression 1 from the coefficients of PRE5 through PRE12 as compared to those of PRE3 and PRE4. The coefficient of the omitted dummy variable PRE2 is, of course, zero.) The gap between these two groups narrows by the term's end. Other things equal, a student who scored 12 or more correct answers on the pre-test can be expected to do only about 5 points better on the post-test than a student who had correctly answered only 4 or fewer questions on the pre-test.

Course grades appear to be fairly well related to student knowledge, even when factors related to student background and instructor leniency are controlled for. This can be seen from the coefficients of KDUM in Regression 2. (KDUM is the dummy version of the KNOW variable. See Table 1 for definitions.) Students who scored less than 8 correct answers on the post-test do significantly worse in the course than

those whose knowledge is greater. At the extremes (cf.: coefficient of KDUM16), the difference in grade can be as great as 20 points.[10]

Instructors and knowledge. The coefficients of INST in Regression 1 give us our measure of instructors' contribution to student knowledge. The numéraire (omitted) instructor is INST1; since all other b_k's are positive, his is the least contribution. The contribution (or value added) of instructors 3, 9, and 11 is not significantly greater than his. At the other end of the scale is instructor 10, whose students can be expected to score three points higher on the post-test than students of instructor 1, even when possible differences in class composition, etc. are controlled for. (A difference of three points on a nineteen-question test is quite substantial; recall that the difference between the overall post-test mean score and the overall pre-test mean was about 4.3 points.)

Instructors and leniency. Instructors appear to differ substantially in their liberality in grading. From the coefficients of the INST variables in Regression 2, we note that INST1, the reference instructor, is the toughest grader. Several instructors are not significantly more lenient than he is. But a student of a given background, with a given level of knowledge of economics, can expect to receive a grade from five to eleven points higher from other instructors.[11]

---

[10] The coefficients of KDUM fall into several groups. Holding other factors constant, post-test scores of 8 - 10 result in a percent grade about 5 points higher than post-test scores below 8. Post-test scores of 11 - 14 are "worth" about 10 - 12 extra percentage points, while post-test scores of 15 or better yield an extra 17 - 20 points in grades.

[11] Some interesting sidelights: The instructor with the greatest value added (INST10) is one of the least lenient, while the instructor with the least value added (INST1) is also one of the least lenient. The most lenient instructor (INST11) has a value added not significantly greater than that of the reference instructor.

The disentangled coefficients of Equation 2A are presented as Regression 2A. While there are some slight differences between the coefficients of Regression 2 and Regression 2A, these appear to be negligible. Even though several of the independent variables are statistically significant in explaining knowledge, the anticipated problem of multicollinearity seems small, perhaps because these variables explain only about 29% of the variation in KNOW.

3.2 Value added, leniency, and evaluations

Having arrived at measures of each instructor's contribution to students' knowledge and his leniency in grading, we are now in a position to confront the central question of this study: To what extent are instructor leniency and "value added" rewarded by high evaluations? Our measure of contribution to knowledge (CONTRIB) is the set of estimated coefficients $\{b_1,\cdots,b_{14}\}$ from Regression 1; our measure of leniency (LEN) is the set of estimated coefficients $\{d_1,\cdots,d_{14}\}$ from Regression 2. When E, the section mean responses to the "overall effectiveness" question, is regressed on these variables plus an intercept term, the result is:

Regression 3A (standard errors in parentheses)

$$E = \underset{(0.465)}{3.37} - \underset{(0.216)}{0.124}\ \text{CONTRIB} - \underset{(0.55)}{0.086}\ \text{LEN}$$

$$R^2 = 0.186$$

Both coefficients are quite close to and not significantly different from zero. Apparently, neither leniency in grading nor contribution to students' knowledge has appreciable influence on what students consider "effective teaching". In order to correct for what may have been a subjective response

by students to instructors with a foreign (i.e., non-North American) accent, we estimated Equation 3, including a dummy variable FOR (whose value is one for instructors whose mother tongue was not English):

Regression 3B (standard errors in parentheses)

$$E = \underset{(.328)}{3.37} - \underset{(.153)}{0.084}\,CONTRIB - \underset{(.043)}{0.020}\,LEN - \underset{(.250)}{0.870}\,FOR$$

$$R^2 = 0.632$$

Although inclusion of FOR substantially improves the fit of the evaluation equation, the impact of CONTRIB and LEN becomes even smaller.[12]

These results, along with various other tests of the robustness of Regressions 3A and 3B,[13] suggest that in evaluating an instructor's "overall effectiveness" students are not primarily (or even strongly) responsive either to the instructor's ability in developing their knowledge of economics or to the severity of the instructor's grading of student performance.

---

[12]The results using the coefficients from Regression 2A rather than Regression 2 are essentially no different.

Regression 4A

$$E = \underset{(0.456)}{3.36} - \underset{(.242)}{0.187}\,CONTRIB - \underset{(0.055)}{0.084}\,LEN$$

$$R^2 = 0.175$$

Regression 4B

$$E = \underset{(0.318)}{3.38} - \underset{(0.170)}{0.103}\,CONTRIB - \underset{(0.042)}{0.025}\,LEN - \underset{(0.242)}{0.860}\,FOR$$

$$R^2 = 0.635$$

[13]Other independent variables which might influence student ratings of instructor effectiveness are the teaching experience and the sex of the instructor. We reran Regressions 3A and 3B with variables accounting for each instructor's total previous teaching experience, previous principles experience, or the square roots of each of these, with no changes in the results reported above. Size of class and the time the class met were also insignificant. We could not include a dummy variable for sex of the instructor because we had only one female instructor in our sample. The results were also unchanged when we dropped INST(1) or INST(11) (both outliers in some sense) or all instructors with a foreign accent from our sample. None of the instructors in our sample was French-Canadian, and none had British, Irish, or Australian accents. There was also no change in the results when we used final course grades to estimate the leniency of the instructors.

4. Concluding Remarks

If, as our results indicate, evaluations do not depend on leniency, why have some other studies found a positive relationship between grades and evaluations? Presumably this observed relationship in these studies is not proxying for a positive relationship between learning and evaluations, since this relationship also was not borne out in our study. Two possibilities immediately come to mind: (1) the students in different studies are not random samples from the entire population of students; (2) in the studies which used individual data instead of section averages, the observed results may be picking up the possibility that those instructors taught primarily to the brighter students (who consequently received higher grades). Such behaviour would have been masked by our use of section averages.

We would like to stress that we have not attempted in this study to capture all of the factors that determine evaluations; we have not attempted, in other words, to estimate the equation that best predicts E. What is being measured by student evaluations of teaching effectiveness remains an open question and a disturbing one. Our findings lead us to believe that students evaluate instructors on the basis of fairly subjective feelings which are not related in any direct way either to the grades they receive or to how much they learn from the instructor. High ratings for "effective teaching" may thus go to instructors who have good rapport with students, who show "concern" for students, or provide a pleasant classroom atmosphere. First year students (who comprise 81.2% of our sample) may be particularly sensitive to instructor characteristics which help make their transition from high school to university less painful. Such characteristics may bear little relationship to leniency in grading or ability to convey knowledge

of the subject.[15]

This kind of student response is consistent with the notion that university attendance is to a large extent a consumption activity. Students rate highly those instructors who provide a high quality of the consumption good. In rewarding instructors with high evaluations, university administrators may not be rewarding the best teachers (if teaching is taken to mean contribution to student knowledge) but are providing incentives for instructors to develop whatever characteristics go into producing the consumption good. It is hard to see how such an incentive system could help build or maintain great universities. In times of sagging enrolments (and the attendant financial crunches), however, the short-run appeal of such a reward structure may be irresistible.

While we place a great deal of confidence in our results, we should emphasize that they have been obtained from one beginning course in one department in one university. The results might be different for a different department, for students taking an upper-level course, or for different types of students at different universities. We strongly suspect that replications of this experiment will yield similar results, but we encourage those interested in pursuing the question further to adopt the approach we have used and to measure learning and leniency as accurately as possible.

---

[15] To the extent that upperclass students have made the adjustment to university, we would expect them to respond somewhat differently. If evaluations were available on an individual student basis tests of this hypothesis would be most interesting.

REFERENCES

D. R. Capozza, "Student Evaluations, Grades, and Learning in Economics," W.E.J. (March, 1973).

F. Costin, W. T. Greenough, and R. J. Menges, "Student Ratings of College Teaching," Rev. of Educ. Res. (1971).

R. W. Crowley and D. A. Wilton, "An Analysis of 'Learning' in Introductory Economics," C.J.E. (Nov., 1974).

K. O. Doyle, "Generalizability of Evaluative Data," presented to the University of Manitoba Symposium on College Teaching and Its Evaluation (Oct., 1974).

_____ and S. E. Whitely, "Student Ratings as Criteria for Effective Teaching," Amer. Ed. Res. Jl. (Summer, 1974).

K. E. Eble, "What are We Afraid Of?" College English (Jan., 1974).

P. W. Frey, "Student Ratings of Teaching: Validity of Several Rating Factors," Science (vol. 182, 1973).

P. K. Gessner, "Evaluation of Instruction," Science (vol. 180, 1973).

A. C. Kelley, "Uses and Abuses of Course Evaluations as Measures of Educational Output," Jl. Ec. Educ. (Fall, 1972).

R. B. McKenzie and G. Tullock, The New World of Economics: Explorations into the Human Experience; Irwin, 1975.

R. Menges, "Evaluating Learning and Teaching," in C. R. Pace (ed.) New Directions for Higher Education (Winter, 1973).

H. G. Murray, "The Validity of Student Ratings of Teaching Ability," presented at the Canadian Psychological Association Meetings (June, 1972).

A. Nichols and J. C. Soper, "Economic Man in the Classroom," J.P.E. (Sept.-Oct., 1972).

R. R. Perry and R. R. Baumann, "Criteria for Evaluation of College Teaching: Their Reliability and Validity at the University of Toledo," in A. L. Sockloff (ed.), Proceedings of the Conference on Faculty Effectiveness as Evaluated by Students; Temple University, 1973.

G. L. Reuber, "Annual Report of the Dean, 1973-74," The University of Western Ontario, 1974.

M. Rodin and B. Rodin, "Student Evaluations of Teachers," Science (vol. 177, 1972).