OPTIMIZING THE FAIRNESS OF STUDENT EVALUATIONS: A STUDY OF CORRELATIONS BETWEEN INSTRUCTOR EXCELLENCE, STUDY PRODUCTION, LEARNING PRODUCTION, AND EXPECTED GRADES

Richard John Stapleton
Gene Murkison
Georgia Southern University

Student evaluations in recent years have become widely accepted as a means of evaluating teachers in higher education. This acceptance was to some extent caused by taxpayers, parents, legislatures, and perhaps students, who demanded more evidence that they were getting good value for their investments (Cone, 1996; England, Hutchings, & McKeachie, 1997; McKeachie, 1997a). This acceptance has not occurred without concerns being voiced by educators about how student evaluations have affected the quality of education. Some worry that the acceptance and use of student evaluations have resulted in grade inflation and lower academic standards. Although many in higher education may have concerns about how specific student evaluation instruments are designed and constructed and how student evaluation numbers are summarized, interpreted, and used, few in higher education today will argue that student evaluations should not be used at all.

At various times during the past 30 years in our department, we experienced conflict caused by student evaluations in the faculty evaluation process. Various faculty members argued that student evaluations were unfair

Authors' Note: Please address correspondence to Richard John Stapleton, Georgia Southern University, P.O. Box 8154, Statesboro, GA 30460-8154; (phone) 912-681-5799; (e-mail) rjstapln@gasou.edu.

JOURNAL OF MANAGEMENT EDUCATION, Vol. 25 No. 3, June 2001 269-291

because students rated some faculty members low as instructors because of the nature and amount of work assigned and the grades students earned (Murkison, 1991; Pickett, 1987; Randall, Price, Tudor, & Stapleton, 1999; Stapleton, 1990; Stapleton & Stapleton, 1996). This article presents findings from the student evaluation literature dealing with general relationships between instructor excellence, study production, learning production, and expected grades production. These relationships are then studied and analyzed using recent departmental data generated by student evaluations.

General Student Evaluation Relationships

Although concerns have been raised by researchers in the student evaluation literature about the validity of student evaluations, the literature indicates that student evaluations are generally valid in the sense that they generally identify relative levels of teaching productivity among teachers as we define teaching productivity, that is, causing students to learn (Cohen, 1981; Gagne, 1977; Marsh, 1980, 1982; Marsh & Roche, 1997).

INSTRUCTOR EXCELLENCE, HOMEWORK, AND LEARNING

The student evaluation literature indicates that in general there are positive correlations between how much students learn in a course and the rating of the instructor, between how much students study and how much they learn in the course, and between the amount and difficulty of work required in the course and the rating of instruction (Greenwald & Gillmore, 1997; Marsh, 1980, 1982; Marsh & Roche, 1997; McKeachie, 1997b). In addition, there are several variables that are significantly positively correlated with ratings of instruction, such as enthusiasm, how clearly one presents the material, whether one answers students' questions, whether one treats students in a courteous and professional manner, whether one returns graded exams and papers promptly, and whether one is well prepared for class (D' Appollonia & Abrami, 1997; Marsh, 1982; Tang, 1997).

EXPECTED GRADES

A great deal of research has been published (Brown, 1976; Cohen, 1981; D' Appollonia & Abrami, 1997; Greenwald, 1997; Greenwald & Gillmore, 1997; Howard & Maxwell, 1980; Kulik & Kulik, 1974; Marchese, 1997; Marsh, 1980, 1981, 1982; Marsh & Roche, 1997; McKeachie, 1997b) showing that there have been concerns in academic environments for years about how grades affect student evaluations. The student evaluation literature indi-

cates that a significant positive correlation does not generally exist between ratings of instructor excellence and final grades assigned by instructors. On the other hand, expected grades have long been suspected of biasing instructor excellence scores (Marsh, 1980).

Greenwald and Gillmore (1997) published new findings about the correlation between instructor excellence scores and expected grades, that is, the final grades students expect to receive in the course at the time they fill out their student evaluation forms. According to Greenwald and Gillmore, there is a significant positive correlation (.45) between expected grades and instructor excellence scores. They also assert that the most important determinant of expected grades bias is not the actual grade a student expects to receive for the course but the relative grade, the grade a student expects relative to the average grade he or she normally receives in courses. Greenwald and Gillmore asserted that a teacher can manipulate higher instructor excellence scores by causing students to expect relatively high grades.

Marsh (1980), Marsh and Roche (1997), and McKeachie (1997b) allowed that in certain cases, it might be possible for a faculty member to lower work requirements or grade leniently to raise grade expectations and perceptions of instructor excellence, but this would not be possible in most cases. They asserted that many students rate as poor instructors who have low standards and who are easy graders. Marsh (1980, 1981) and Marsh and Roche concluded that much of the expected grades bias on student evaluations of instructor excellence is caused by the prior interest of the student in the subject matter taught by the course. McKeachie (1997b) asserted that because students learn more in courses taught by excellent instructors, they would naturally expect higher grades in courses taught by excellent instructors, and Greenwald and Gillmore (1997) did not prove that high expected grades generally bias upward the ratings of the instructor.

Method of This Research

We are concerned in this research with whether student evaluations are fair to all faculty members. Although student evaluations may be generally valid statistically, this does not prove that each student evaluation conducted by every school will be valid in the case of every faculty member included. There is some risk that administrators will rank instructor excellence scores to determine teaching excellence and use these ranks to partially determine salary, tenure, and promotion decisions. McKeachie (1997b) called this practice deplorable.

The first 11 questions are answered with
1 = *strongly agree*, 2 = *agree*, 3 = *neither disagree nor agree*, 4 = *disagree*, and 5 = *strongly disagree*.

1. Overall, the instructor is an excellent teacher.
2. The instructor motivates me to do my best work.
3. The instructor showed genuine concern for the student.
4. The instructor seems well prepared for each class.
5. The instructor is enthusiastic about the subject matter.
6. I would recommend the instructor to a friend.
7. The instructor presented the material clearly and effectively.
8. The instructor evaluates in a fair manner.
9. I usually give lower ratings to instructors who require a lot of work.
10. The instructor is timely in providing feedback on my work.
11. I think that courses that require a lot of work are more valuable than courses that don't.
12. On average, the number of hours I studied per week outside of class for this course was
    1 = *less than 1 hour*, 2 = *1-3 hours*, 3 = *4-6 hours*, 4 = *7-9 hours*, 5 = *10 hours or more*.
13. Compared to other courses I have taken, in this course, I have learned
    1 = *much more*, 2 = *more*, 3 = *no more or less*, 4 = *less*, 5 = *much less*.
14. Given my efforts in this course, the grade I expect to receive may not be the same I think I deserve. It will be
    1 = *much lower*, 2 = *lower*, 3 = *the same*, 4 = *higher*, 5 = *much higher.*
15. The primary reason I signed up for this course is
    1 = *I like the prof's teaching style*, 2 = *required and only section available*, 3 = *prof recommended by friend*, 4 = *subject of interest to me*, 5 = *I thought easy to make good grade*, 6 = *none of the above*.

Figure 1: The Fall 1996 Department Student Evaluation Form

The major purpose of this research was to determine to what extent there were exceptions in the case of our department to the general relationships generated by the student evaluation research literature in regard to relationships between ratings for instructor excellence and study production, learning production, and expected grades production. The faculty of the department were asked to participate in this study during the fall quarter of 1996.

The department added student study, learning, and relative expected grades questions to the departmental evaluation form in the fall of 1996. A copy of this form is presented in Figure 1. A total of 29 faculty members agreed to the inclusion of questions 9 through 15 with the responses to be analyzed for research purposes only. Fifty-four classes completed the evaluations for a total of 1,251 survey instruments. The first 8 questions on the form were used campuswide, and the remaining 7 were added by the department. Department faculty voted to include the student work and learning questions for research purposes only. The researchers were furnished with anonymous

faculty data by administrators of the department, and the research was approved by the university.

The eight campuswide questions were designed to measure student opinion regarding instructor excellence. The factors used to define instructor excellence include ability to motivate the students to do their best work, level of enthusiasm for the subject matter, ability to convey material clearly, fairness in evaluation, and preparedness for class. The opinion measure considered by most administrators and faculty to be the summarizing component, however, was the first item, "Overall the instructor is an excellent teacher." This question was used as the instructor excellence variable in this study.

The relative frequency, median, and mean of responses to each question, by faculty member and for the entire department, were computed. The analysis prepared for each faculty member included a summary of the results for the department, the individual's results, a graphical comparison of the relative frequency results for the faculty member versus the department, and a chi-square goodness-of-fit analysis for each of the 15 questions. In addition, each faculty member received a typed report of all student comments. It was observed that students were more likely to write comments when evaluations were conducted at the beginning of a class. Therefore, the evaluations were administered at the beginning of a class meeting of each course section during the last few days of the term. No faculty member administered his or her own evaluation process; rather, another faculty member, a graduate assistant, or a staff person conducted the process.

In this study, faculty members are ranked to illustrate how rank order changes depending on which teaching/learning variables are considered. Pearson's and Spearman's correlations were computed to test for relationships among faculty means and rankings in regard to instructor excellence, study production, learning production, and expected grades production. On the other hand, correlation coefficients between all questions were calculated using data aggregated by the department as a whole. SPSS was employed to calculate the correlation matrix for the means and medians of the first 14 questions for the 54 sections of class taught by departmental faculty. Since the same relationships were observed using medians and means, the medians were eliminated from further study. Five hypotheses were developed.

Hypotheses

STUDY PRODUCTION

Murkison (1991) was concerned about the influence of student workloads on student evaluations. He found that whereas some instructors who required

a great amount of outside study were rated highly, others who did not require a high level of study were also rated highly. These bipolar results confound a standard linear statistical analysis. On the other hand, Murkison concluded that outside study loads were negatively correlated with student evaluations when the class was composed of many students who were initially misinformed about the quantity of work required. Based on this research and our observations over the years, we would expect to find a nonsignificant linear statistical relationship here due to the bipolar nature of the findings. This leads to Hypothesis 1.

Hypothesis 1: Instructor excellence (Question 1) will be negatively correlated with study production (Question 12).

LEARNING PRODUCTION

Stapleton and Stapleton (1996, 1998) reported on student evaluations of one professor in the department and learning effectiveness data gathered by forms administered individually by the same professor. This study examined evaluations for this professor who, since 1994, had been ranked as relatively low in excellent instructor rankings but had been ranked highly by the same students when asked how much they had studied and learned in the course. This study also included data showing this professor had been rated highly in terms of study and learning production in research spanning 15 years. This leads to our second hypothesis.

Hypothesis 2: Instructor excellence (Question 1) will be positively correlated with learning production (Question 13).

GRADE EXPECTATIONS

As shown in the literature (Greenwald & Gillmore, 1997; Marsh, 1980, 1981, 1982; Marsh & Roche, 1997), expected grades have long been suspected of biasing the evaluation of the instructor. With this hypothesis, we tested whether in our department there was a positive relationship between relative expected grades, as emphasized by Greenwald and Gillmore, and the instructor excellence rating.

Hypothesis 3: Instructor excellence (Question 1) will be positively correlated with expected grades production (Question 14).

STUDY AND LEARNING

The correlation analyses by Pickett (1987), Murkison (1991), and Stapleton and Stapleton (1996) consistently showed positive relationships between perceived student learning and reported hours spent studying and/or preparing outside of class. Our efforts with this hypothesis were designed to replicate these findings. Accordingly, we expected to find a positive relationship in this research.

Hypothesis 4: Study production (Question 12) will be positively correlated with learning production (Question 13).

CONFOUNDING RELATIONSHIPS

All of the referenced research at our institution (Murkison, 1991; Pickett, 1987; Randall et al., 1999; Stapleton & Stapleton, 1996) found that some professors rank in opposition to the norm or common relationships. That is, they may rank high as excellent instructors (Question 1) but low in study production (Question 12), learning production (Question 13), and/or expected grade production (Question 14). Conversely, some professors will rank relatively low as excellent instructors but rank higher in terms of study, learning, and/or expected grades. We have heard these sentiments expressed many times within the department.

Hypothesis 5: Some professors will rank relatively high in instructor excellence (Question 1) but relatively low in learning production (Question 13), and some professors will rank relatively low in instructor excellence but high in learning production.

Results

As was expected and as shown in Table 1, all the instructor excellence variables measured with Questions 2 through 8 were strongly, positively correlated with Question 1, the overall excellence item. Following is a discussion of the findings in relation to the hypotheses.

HYPOTHESIS 1

Hypothesis 1 was not substantiated in Table 1, which aggregates the data by the department as a whole, because there was not a statistically significant negative correlation between the instructor excellence rating (Question 1)

TABLE 1
Correlation Coefficients for Means of All Responses to All Questions

	Excellent Instructor	*Motivation*	*Concerned*	*Prepared*	*Enthusiastic*	*Recommend*
Excellent instructor		.8558***	.6996***	.6074***	.6128***	.8416***
Motivates	.8558***		.6342***	.5976***	.6021***	.7285***
Concerned	.6996***	.6342***		.4966***	.6098***	.7729***
Prepared	.6074***	.5976***	.4966**		.4726***	.6001***
Enthusiastic	.6128***	.6021***	.6098***	.4726***		.5220***
Recommend	.8416***	.7285***	.7729***	.6001***	.5220***	
Clear	.8587***	.6789***	.6883***	.7531***	.5013***	.8359***
Fair	.2953**	.3567***	.2407*	.1245	.1143	.2179
Rate lower when much work	.3735***	.4170***	.2029	.202	.2831**	.3663***
Timely feedback	.4010***	.4319***	.4925***	.4489***	.3478**	.3911***
Work value	.1007	.2211	.0355	.0681	.0294	.0275
Hours	–.0938	–.1581	–.0404	.0178	.0033	.1574
Learned	.7500***	.7636***	.4697***	.5720***	.4429***	.5882***
Grade	–.2604*	–.2933**	–.2938**	–.1773	–.1649	–.3590***

* $p \leq .10$. ** $p \leq .05$. *** $p \leq .01$.

and the ratings for study production (Question 12). Due to the direction of ordinal scaling, the negative correlation (–.0938) shown in Table 1 between Questions 1 and 12 reflects a nonsignificant positive relationship.

This positive relationship, however, was not shown in Table 2, which compares the means for specific faculty members (Pearson's correlation = .1449, $p \leq .45$). A nonsignificant negative relationship is shown between instructor excellence and study production due to the ordinal favorable and unfavorable reversing of the question responses.

Comparing ranks in Table 2, however, this hypothesis was substantiated because the Spearman's rank correlation coefficient was –.2793, $p \leq .07$, indicating a statistically significant negative correlation between study production and instructor excellence because the ranks for both questions had the same direction of ordinal scaling.

HYPOTHESIS 2

Hypothesis 2 was substantiated. A strong positive correlation (.7500) ($p \leq .01$) was observed between instructor excellence (Question 1) and learning production (Question 13) in Table 1. This positive relationship between

Clear	*Fair*	*Rate Lower When Much Work*	*Timely Feedback*	*Work Value*	*Hours*	*Learned*	*Grade*
.8587***	.2955**	.3735***	.4010***	.1007	–.0938	.7500***	–.2604*
.6789***	.3567***	.4170***	.4319***	.2211	–.1581	.6360***	–.2933**
.6883***	.2407*	.2029	.4925***	.0355	–.0404	.4697***	–.2938**
.7531***	.1245	.202	.4483***	.0681	.0178	.5720***	–.1773
.5013***	.1143	.2831**	.3478**	.0298	.0033	.4429***	–.1648
.8359***	.2179	.3663***	.3911***	.0275	.1574	.5882***	–.3590***
.2283	.2615*	.3940***	.0367	–.0149	.6803***	–.3237**	
.2283		.0757	.0861	–.1710	–.0248	.2887**	.0268
.2615*	.0757		.0721	–.1964	.2804**	.151	.0079
.3940***	.0861	.0721		.2604*	–.1193	.3801***	–.1399
.0367	–.1710	–.1964	.2604*		–.5367***	.4021***	.0464
–.0149	–.0248	–.2804**	–.1193	–.5367		–.3466**	–.1049
.6803***	.2887**	.151	.3801***	.4021***	–.3466**		–.2572*
–.3237**	.0268	.0079	–.1399	.0464	–.1049	–.2572*	

Questions 1 and 13 is also clearly shown in Table 3, which compares means and ranks by department member.

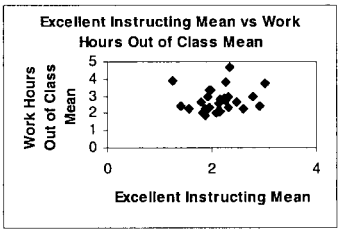
HYPOTHESIS 3

Hypothesis 3 was substantiated. The higher the instructor excellence score (Question 1), the higher the expected grade in the course (Question 14), and vice versa. This is shown by the –.2604 ($p \leq .1000$) correlation between Questions 1 and 14 in Table 1 because the favorable and unfavorable directions of responses for the questions generating the means are reversed. This is a positive relationship between instructor excellence and expected grades. This correlation is shown more robustly in Table 4 in terms of means (–0.5662) and ranks (–0.6142), where negative statistical correlations indicate a positive relationship between expected grades and instructor excellence.

Faculty members with low grade expectation means were assigned better ranking numbers than faculty with high grade expectation means, that is, the lower the mean and the lower the actual expected grades, the closer the rank was to first place. The scatter diagram of ranks in Table 4 shows this relationship visually.

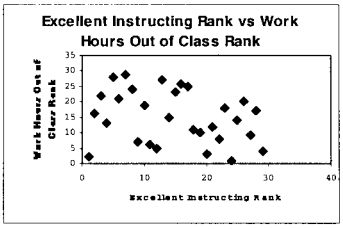
TABLE 2
Instructor Excellence Versus Study Production

Faculty Member	Instructor Excellence	Work Hours Out of Class		
		Rank	*Mean*	*Rank*
R	1.25	1	3.94	2
N	1.4	2	2.45	16
U	1.55	3	2.25	22
I	1.79	4	2.68	13
X	1.82	5	2	28
Q	1.86	6	2.24	21
G	1.87	7	1.85	29
H	1.89	8	2.15	24
L	1.91	9	2.98	7
P	1.94	10	2.35	19
Al	1.95	11	3.37	6
O	1.97	12	3.38	5
W	2.07	13	2.02	27
Bl	2.12	14	2.57	15
Cl	2.12	15	2.19	23
T	2.13	16	2.13	26
E	2.15	17	2.11	25
V	2.15	18	2.78	11
Bl	2.23	19	2.91	10
Z	2.25	20	3.82	3
Y	2.26	21	2.75	12
Cl	2.3	22	2.95	8
J	2.31	23	2.35	18
S	2.33	24	4.67	1
D	2.46	25	2.67	14
F	2.59	26	2.3	20
M	2.79	27	2.93	9
K	2.91	28	2.43	17
A	3	29	3.78	4



Pearson Correlation Coefficient

Coefficient of Correlation	0.14494
t statistic	0.76119
p-value	0.45314



Spearman Correlation Coefficient

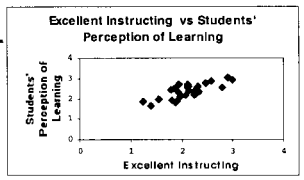
Coefficient of Correlation	-0.27931
t statistic	-1.51150
p-value	0.07114

HYPOTHESIS 4

Hypothesis 4 was substantiated in Table 1, which aggregates data by the whole department, because the relationship between study production (Question 12) and learning production (Question 13) was significantly positive. This is indicated by a negative (–.3466) ($p \leq .05$) statistical correlation due to the ordinal scaling. On the other hand, as shown in Table 5, which compares means and ranks between faculty members, the hypothesis was not substantiated because there is little correlation between study production and learning production.

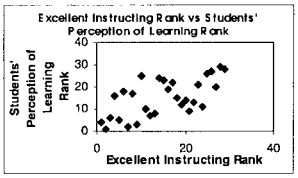
TABLE 3
Instructor Excellence Versus Learning Production

Faculty Member	Excellent Instructing		Students' Perception of Their Own Learning	
	Mean	*Rank*	*Mean*	*Rank*
R	1.25	1	1.88	4
N	1.4	2	1.65	1
U	1.55	3	1.95	6
I	1.79	4	2.45	16
X	1.82	5	1.94	5
Q	1.86	6	2.5	18
G	1.87	7	1.8	2
H	1.89	8	2.49	17
L	1.91	9	1.85	3
P	1.94	10	2.74	25
AI	1.95	11	2.29	10
O	1.97	12	2.08	7
W	2.07	13	2.21	8
BI	2.12	14	2.73	24
CI	2.12	15	2.64	23
T	2.13	16	2.56	19
E	2.15	17	2.64	22
V	2.15	18	2.41	15
BI	2.23	19	2.36	12
Z	2.25	20	2.38	14
Y	2.26	21	2.21	9
CI	2.3	22	2.37	13
J	2.31	23	2.61	21
S	2.33	24	2.33	11
D	2.46	25	2.77	26
F	2.59	26	2.89	27
M	2.79	27	2.57	20
K	2.91	28	3.05	29
A	3	29	2.95	28



Pearson Correlation Coefficient

Coefficient of Correlation	0.75871
t statistic	6.05192
p-value	0.00000



Spearman Correlation Coefficient

Coefficient of Correlation	0.63842
t statistic	4.31000
p-value	0.00010

HYPOTHESIS 5

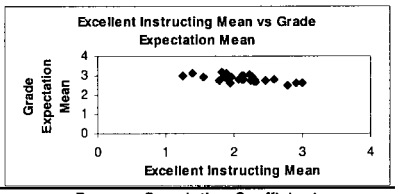
Hypothesis 5 was substantiated. As shown in Table 6, which summarizes the ranks in Tables 2 through 4, it can be seen that four faculty members achieved ranks that were significantly better for instructor excellence (Question 1) than ranks they achieved for learning production (Question 13), and four faculty members achieved ranks for learning production (Question 13) that were significantly better than their ranks for instructor excellence (Question 1). We considered a rank significantly better for a faculty member if it resulted in the faculty member's being ranked as a teacher in the upper half of the department rather than in the lower half. A summary of the findings is presented in Figure 2.

Analysis and Conclusions

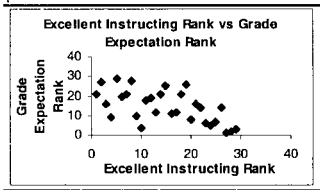
Our results replicated findings in the student evaluation literature that student evaluations are generally valid by showing a positive relationship between

TABLE 4
Instructor Excellence Versus Expected Grades

Faculty Member	Excellent Instructing		Grade Expectation	
	Mean	*Rank*	*Mean*	*Rank*
R	1.25	1	3	21
N	1.4	2	3.1	27
U	1.55	3	2.93	16
I	1.79	4	2.78	9
X	1.82	5	3.18	29
Q	1.86	6	2.96	20
G	1.87	7	3	21
H	1.89	8	3.11	28
L	1.91	9	2.8	10
P	1.94	10	2.65	4
AI	1.95	11	2.95	18
O	1.97	12	2.95	19
W	2.07	13	2.82	12
BI	2.12	14	3	21
CI	2.12	15	3.02	25
T	2.13	16	2.81	11
E	2.15	17	2.82	12
V	2.15	18	3	21
BI	2.23	19	3.09	26
Z	2.25	20	2.75	8
Y	2.26	21	2.92	16
CI	2.3	22	2.84	14
J	2.31	23	2.69	6
S	2.33	24	2.67	5
D	2.46	25	2.74	7
F	2.59	26	2.84	14
M	2.79	27	2.5	1
K	2.91	28	2.61	2
A	3	29	2.6	3



Pearson Correlation Coefficient	
Coefficient of Correlation	-0.566256158
t statistic	-3.569823654
p-value	0.000682210



Spearman Correlation Coefficient	
Coefficient of Correlation	-0.61429
t statistic	-4.04510
p-value	0.00021

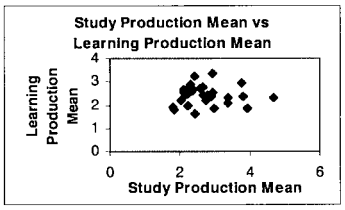
instructor excellence scores and learning produced in the course and that students expect high grades in courses taught by teachers they rate highly as instructors. Our findings, however, did not fully replicate the findings of the student evaluation literature regarding the relationship between instructor excellence and study production or the relationship between study production and learning production. We also found several faculty members who confounded the general relationship between instructor excellence and learning production.

CONFLICTING RESULTS

According to the student evaluation literature, students do not generally rate as poor teachers who assign more homework than others, and students generally learn more the more they study. As shown in Table 2, based on the Spearman's rank correlation, we found a statistically significant negative correlation between instructor excellence and study production, meaning the teachers who assigned more homework were in general rated lower as

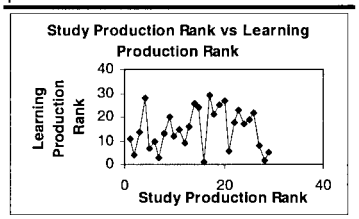
TABLE 5
Study Production Versus Learning Production

Faculty Member	Study Production		Learning Production	
	Mean	*Rank*	*Mean*	*Rank*
R	3.94	2	1.88	4
N	2.45	16	1.65	1
U	2.25	22	1.95	6
I	2.68	13	2.45	16
X	2	28	1.94	5
Q	2.24	21	2.5	18
G	1.85	29	1.8	2
H	2.15	24	2.49	17
L	2.98	7	1.85	3
P	2.35	19	2.74	25
AI	3.37	6	2.29	10
O	3.38	5	2.08	7
W	2.02	27	2.21	8
BI	2.57	15	2.73	24
CI	2.19	23	2.64	23
T	2.13	26	2.56	19
E	2.11	25	2.64	22
V	2.78	11	2.41	15
BI	2.91	10	2.36	12
Z	3.82	3	2.38	14
Y	2.75	12	2.21	9
CI	2.95	8	2.37	13
J	2.35	18	2.61	21
S	4.67	1	2.33	11
D	2.67	14	2.77	26
F	2.3	20	2.89	27
M	2.93	9	2.57	20
K	2.43	17	3.05	29
A	3.78	4	2.95	28



Pearson Correlation Coefficient

Coefficient of Correlation	-0.00795
t statistic	-0.04131
p-value	0.48367



Spearman Correlation Coefficient

Coefficient of Correlation	0.09754
t statistic	0.49973
p-value	0.31074

instructors. As shown in Table 5, also based on a Spearman's rank correlation, we found little correlation between study production and learning production, meaning students did not necessarily think they learned more in courses in which they studied more.

The scatter diagram for ranks in Table 5 shows a generally concave-from-below curve wherein teachers receiving the lowest and the highest study production ratings generally received the highest learning production ratings, whereas teachers in the middle of the study production rankings were generally rated lower in learning production. A relevant consideration is whether students are estimating their learning relative to the total learning required by the teacher or whether they are they estimating how much they learned from the teacher relative to how much they normally learn from teachers. Some students may think they learn a great deal from teachers who require little homework because they are able to retain and understand a high per-

TABLE 6
Rankings of Relevant Variables

	Rank			
Faculty Member	*Instructor Excellence*	*Study Production*	*Learning Production*	*Expected Grades Production*
R	1	2	4	21
N	2	16	1	27
U	3	22	6	16
I	4	13	16	9
X	5	28	5	29
Q	6	21	18	20
G	7	29	2	21
H	8	24	17	28
L	9	7	3	10
P	10	19	25	4
AI	11	6	10	18
O	12	5	7	19
W	13	27	8	12
BI	14	15	24	21
CI	15	23	23	25
T	16	26	19	11
E	17	25	22	12
V	18	11	15	21
B	19	10	12	26
Z	20	3	14	8
Y	21	12	9	16
C	22	8	13	16
J	23	18	21	6
S	24	1	11	5
D	25	14	26	7
F	26	20	27	14
M	27	9	20	1
K	28	17	29	2
A	29	4	28	3

centage of what is presented, whereas some students may think they learn a great deal from teachers who require much homework because they learn more in those courses than they normally learn in courses due to the broader or more rigorous coverage or content. Thus, the quality and quantity of learning produced by teachers with similar learning production ratings may not be the same.

Truth—with respect to relationships between instructor excellence, study production, and learning production—is relative to the point of view of the

Hypothesis	*Finding*
Hypothesis 1	
Instructor excellence negatively related to study production	Not substantiated for whole department (see Table 1) Substantiated based on faculty ranks (see Table 2)
Hypothesis 2	
Instructor excellence positively related to learning production	Substantiated (see Tables 1 and 3)
Hypothesis 3	
Instructor excellence positively related to expected grades production	Substantiated (see Tables 1 and 4)
Hypothesis 4	
Study production positively related to learning production	Substantiated for whole department (see Table 1) Not substantiated based on faculty ranks (see Table 5)
Hypothesis 5	
Some teachers rank high in instructor excellence but low in learning production and vice versa	Substantiated (see Table 6)

Figure 2: Summary of the Findings

observer. Viewing the student evaluation problem by focusing on students en masse, as in Table 1, it is generally true that the more students study, the more they learn and the higher they rate the teacher. On the other hand, viewing the problem by focusing on specific teachers and computing correlations from average ratings for study production and learning production, as in Table 5, it is not generally true that the more students study, the more they learn. Nor is it true, when one focuses on specific teachers, as in Table 2, that the more students study, the higher they rate the teacher.

This conundrum is caused by the fact that study production, learning production, and instructor excellence levels among teachers are different, and these phenomena are hidden when one aggregates data for correlation studies by students, as in Table 1. When data are aggregated for correlation studies by students en masse, the questionnaires are read by the computer as though the students doing the faculty evaluation were taking a huge course taught by a single teacher. Although some of the correlation studies in the student evaluation literature have aggregated data by courses, many correlation studies in the student evaluation literature have aggregated data by students. Our study

may be the first student evaluation correlation study to aggregate data by specific teachers.

EXCEPTIONS TO THE RULE

As shown in Table 6, the general proposition that a significantly positive relationship exists between instructor excellence and learning production, as asserted in the student evaluation literature and as replicated in our research (Tables 1 and 3), was confounded by 8 of 29 faculty members in our study (29%) because 4 faculty members (I, Q, H, and P) ranked relatively high as excellent instructors but relatively low as producers of learning, and 4 faculty members (S, C, Y, and B) ranked relatively low as excellent instructors but relatively high as producers of learning. These findings are significant because they prove that one cannot logically deduce, using the above proposition, that a faculty member scoring poorly in instructor excellence will also score poorly in learning production or that a faculty member scoring well in instructor excellence will also score well in learning production. It is unlikely that our case is one of a kind, and these results can probably be generally replicated in many educational environments.

FAIRNESS AND ETHICS

From the perspective of specific teachers seeking fairness and justice in faculty evaluations, general student evaluation correlations may be red herrings. The relevant consideration, if fairness is to be provided for all teachers, is not whether there is a positive correlation between instructor excellence and learning production in general or between any two student evaluation variables in general but how each teacher ranks in terms of instructor excellence, study production, learning production, and relative expected grades production in specific student evaluations.

Although instructor excellence is normally weighted and ranked after student evaluations, study production, learning production, and expected grades often are not weighted and ranked because many student evaluation forms do not include study production, learning production, and relative expected grades questions. This automatically creates unfairness.

Unfairness is created when teachers are not recognized and rewarded for the value they produce in the classroom. This can happen when teachers teach subjects or use teaching methods that naturally require more homework than most other courses or methods for minimal levels of student understanding and mastery to occur. In such cases, if only instructor excellence is weighted and ranked, the conscientious teacher will probably receive no recognition for the study and learning he or she produced, and adding insult to

injury, he or she may be forced to suffer the indignity of being ranked low as an instructor due to doing what his or her subject or teaching method requires. Weighting and ranking study production and learning production will eliminate the possibility of occurrence of these inequities and indignities in worst-case forms.

One can argue that it is unfair to teachers who do not produce high expected grades if other teachers in the same student evaluation are allowed to increase their expected grades with impunity. Ranking expected grades will help control teachers who might be tempted to increase expected grades by decreasing homework and lowering grading standards to increase their instructor excellence scores. In our opinion, although this tactic may not occur in a large percentage of cases, it will occur in some percentage of cases. Greenwald and Gillmore (1997) asserted with good evidence and analysis that this tactic may occur in many cases, whereas Marsh and Roche (1997) and McKeachie (1997b) asserted that Greenwald and Gillmore's conclusions are overstated.

In our opinion, neither the student evaluation literature nor this research has proved which is the causative variable in most cases—high instructor excellence ratings or high relative expected grades. On the other hand, although it has not been proven which is generally the chicken or the egg, we know these variables are significantly positively correlated, and our data support the proposition that high expected grades will bias upward instructor excellence scores in some cases.

We also know, based on ranks in Table 2, that instructor excellence and study production may be significantly negatively related in some student evaluations. Although it cannot be proven whether this negative relationship occurred in our study because some faculty deliberately lowered their homework requirements to produce higher expected grades and instructor excellence scores or whether this naturally and ethically occurred because of the different teaching styles and methods of teachers or the difficulty of the subjects taught, the existence of this relationship in this study strongly indicates that it is possible to increase instructor excellence scores by decreasing homework and lower instructor excellence scores by increasing homework, all other things being equal. This possibility creates temptations, conflict, and pitfalls for teachers who are worried about instructor excellence scores' affecting their tenure, merit raises, or reputations.

Regardless of the cause-effect relationships among the variables, it can be seen by scanning the ranks of teachers in Table 6 that student evaluation forms weighting only instructor excellence would create unfairness for some teachers and overvalue others if resulting instructor excellence ratings were assumed to be congruent with the teaching productivity of all included teach-

ers. All one has to do is observe the inverse relationships between instructor excellence and learning production for various teachers in Table 6 to understand the unfairness that would exist if a learning production question were not included on the student evaluation form.

A related ethics issue is whether it is fair to imply that a faculty member who scores relatively high in instructor excellence and relatively low in learning production is a better teacher than a faculty member who scores relatively low in instructor excellence and relatively high in learning production, as is implied if instructor excellence is deliberately ranked in isolation. Among the other combinations and possibilities in Table 6, is a faculty member who scores high in instructor excellence, high in study production, and high in learning production and who causes students to expect high grades a better or worse teacher than a faculty member who scores high in instructor excellence, high in study production, and high in learning production but causes students to expect low grades? One could build the case that the latter teacher may be the better teacher because there is no probability his or her high instructor excellence scores were partially caused by high expected grades.

Weighting and ranking study production, learning production, and relative expected grades production will also help control administrators who might be tempted to psychologically encourage faculty to lower homework requirements and grading standards to improve perceptions of instructor excellence in their administrative units or institutions in the eyes of financial supporters and student customers. Although it is difficult to prove that psychological messages such as these exist or estimate how pervasive they are in educational environments, we are convinced we have sensed them in a few instances in the past 30 years.

Faculty and administrator committees that deliberately exclude study production, learning production, and relative expected grades questions from student evaluation forms are tacitly condoning the creation of the inequities and indignities discussed above if they are aware of the consequences of their actions. If these committees are aware of the consequences of choosing or excluding particular student evaluation questions, it is possible they may decide that excluding study production, learning production, and relative expected grades questions from the student evaluation form will result in creating the greatest happiness for the greatest number, and thus their decisions may be considered ethical. If they are guilty of anything, it may be that they are simply too human. Unfortunately, whether student evaluation committees were aware of these consequences when they created their student evaluation forms cannot be proven because committee structures eliminate personal accountability.

THE COMPOSITE INDICATOR OF TEACHING PRODUCTIVITY (CITP)

As McKeachie (1997b) and Marsh and Roche (1997) have pointed out, most of the injustices in teaching evaluations are not caused by student evaluations per se but by how the data they generate are used, and both advocate using more than one dimension of teaching to evaluate teachers. They point out weighting only one dimension of teaching favors certain teaching styles and methods and tends to encourage stereotypical behavior among teachers. It seems weighting only instructor excellence would favor lecturing behaviors, whereas also weighting study production, learning production, and expected grades production would lend support for behaviors required to teach discussion, experiential, clinical, and Internet courses.

Our CITP in Table 7 weights equally ranks for instructor excellence, study production, learning production, and expected grades production. A faculty member's CITP score is the average of his or her weighted ranks in the four areas. Different combinations of variables and weights could be used in a CITP in different educational environments. As shown in Table 7, the rank for instructor excellence and the CITP rank are significantly positively correlated (.3594) in this evaluation, meaning the relative evaluations of most faculty members were not significantly affected by the CITP computed in this manner, indicating for most faculty members that using a CITP will not raise a question of fairness.

On the other hand, one might argue that the CITP is unfair because it harms the reputations of teachers who produce high instructor excellence scores, who produce high study and learning, and who naturally and ethically cause their students to expect high grades as a reward for achievements. The counterargument is that not using a CITP would be unfair to teachers who score low in instructor excellence, who produce high study and learning, and who cause their students to expect low grades.

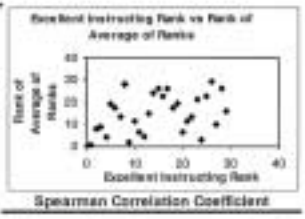
It can be argued, based on utilitarian criteria for justice and fairness (Rawls, 1999), that the greater good for the greater number will be achieved using a CITP because using a CITP will prevent serious harm to the reputations and self-esteem of teachers who produce low instructor excellence scores, high learning scores, and various study and expected grades scores, whereas using a CITP will cause only minor harm to the reputations and self-esteem of teachers who produce high instructor excellence scores, high study scores, high learning scores, and high expected grades.

USING STUDENT EVALUATION RESULTS

Some argue that computers and statistics should not be used to compute even means and medians from student evaluations to facilitate comparisons

TABLE 7
Composite Indicator of Teaching Productivity

Faculty Member	Instructor Excellence		Composite Indicator (CITP)	
	Mean	Rank	Average of Ranks	Rank of Rank Averages
R	1.25	1	7.00	1
N	1.40	2	11.50	8
U	1.55	3	11.75	9
I	1.79	4	10.50	4.5
X	1.82	5	10.75	10.5
Q	1.86	6	16.25	17.5
G	1.87	7	14.75	13.5
H	1.89	8	19.25	28
L	1.91	9	7.25	2
P	1.94	10	14.50	11.5
AI	1.95	11	11.25	6.5
O	1.97	12	10.50	4.5
W	2.07	13	15.00	15
BI	2.12	14	18.50	24
CI	2.12	15	19.00	26
T	2.13	16	18.00	22.5
E	2.15	17	19.00	26
V	2.15	18	16.25	17.5
B	2.23	19	16.75	19.5
Z	2.25	20	11.25	6.5
Y	2.26	21	14.50	11.5
C	2.30	22	14.75	13.5
J	2.31	23	17.00	21
S	2.33	24	10.25	3
D	2.46	25	18.00	22.5
F	2.59	26	21.75	29
M	2.79	27	14.25	10
K	2.91	28	19.00	26
A	3.00	29	16.00	16



Spearman Correlation Coefficient

Coefficient of Correlation	0.39948
t statistic	2.00174
p-value	0.02773

among teachers for faculty evaluations, much less compute departmental norms for specific questions or CITPs. A faculty member and his or her administrator using a minimally quantitative approach such as this would peruse questions on student evaluation forms to become generally aware of what students indicated for certain evaluative considerations and then estimate an overall teaching rating.

There are advantages and disadvantages to such a system. One advantage is that there would be no risk of a department member's reputation and self-esteem being harmed by his or her student evaluation scores' being known and compared throughout the department, school, college, or university. One disadvantage is that because a department member would have no data to show how he or she produced on a relative basis within the department in the area of teaching, he or she would be powerless to argue with any degree of certainty what his or her merit raise should be, assuming the school has merit raises and assuming teaching is weighted heavily with research and service in

the overall faculty evaluation process. Not providing relative quantitative teaching productivity data in such a case would put a faculty member dealing with an administrator for merit raises in a position roughly analogous to that of a child dealing with a parent for an increase in allowance money.

RECOMMENDATIONS

On balance, we recommend that a CITP such as that in Table 7 be widely used as one indicator among others, such as syllabi, exams, and in-class visits, to increase the fairness of the teaching evaluation process. This necessarily entails including study production, learning production, and relative expected grades questions on all student evaluation forms. Although not computerizing student evaluations and not computing means, medians, norms, chi-square differences, CITPs, and other quantitative indicators might be the best practice in a perfect world, such a practice is unlikely to occur in our academic world or in most academic worlds. Computers and statistics, like student evaluations, are permanent parts of the teaching evaluation process in most universities.

Based on our experience, student evaluation numbers are inevitably used for comparison purposes among faculty members, creating de facto ranks even if the department chairperson does not record ranks. Each faculty member knows his or her student evaluation scores; student evaluation scores are used by faculty to satisfy requirements for tenure, promotion, and posttenure reviews; these reviews are conducted by committees of faculty members; faculty talk and benchmark among themselves using student evaluation scores; student evaluation scores were even published by our student government association 2 years ago because our state has an open records law.

Because it is difficult, if not impossible, to prevent student evaluation scores from being used for comparison purposes among faculty, one can argue the student evaluation process should be made as transparent as possible—by ranking the relevant variables and empowering each faculty member with as much information as possible about her or his relative teaching productivity. A CITP will ensure that the strengths and weaknesses of each faculty member's teaching style and methods, as a matter of reliable and dependable procedure, will be taken into account every time a student evaluation is conducted.

Because a positive linear relationship between study production and learning production ranks does not exist in this study (see Table 5) and because there is a negative relationship between study production and instructor excellence ranks (see Table 2), it is possible for some percentage of faculty members to lower homework requirements and grading standards to increase

expected grades production (see Table 4) and to increase their instructor excellence scores and learning production scores (see Table 3) on some student evaluations; and conversely, it is possible for some percentage of faculty members to lower their instructor excellence scores on some student evaluations by increasing homework requirements, raising grading standards, and lowering expected grades.

Consequently, we recommend that instructor excellence, study production, learning production, and expected grades production be weighted and ranked every time a student evaluation occurs. This will ensure that a modicum of fairness will exist for all teachers subjected to the student evaluation, regardless of the teaching style, teaching method, and ethical system that a teacher might adopt or create to produce learning in students and to satisfy the ego and survival needs and requirements of the teacher, students, colleagues, administrators, and external supporters.

References

Brown, D. L. (1976). Faculty ratings and student grades: A university-wide multiple regression analysis. *Journal of Educational Psychology*, *68*, 573-578.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, *51*, 281-309.

Cone, J. D. (1996, April 5). Letters to the editor. *The San Diego Union-Tribune*, p. B11.

D' Appollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, *52*, 1198-1208.

England, J., Hutchings, P., & McKeachie, W. J. (1997). The professional evaluation of teaching (Occasional Paper No. 33). New York: American Council of Learned Societies.

Gagne, R. M. (1977). *The conditions of learning*. New York: Holt, Rinehart & Winston.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, *52*, 1182-1186.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, *52*, 1209-1216.

Howard, G. S., & Maxwell, S. E. (1980). Correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology*, *72*, 810-820.

Kulik, J. A., & Kulik, C. C. (1974). Student ratings of instruction. *Teaching of Psychology*, *1*(2), 51-57.

Marchese, T. (1997, September/October). Student evaluations of teaching. *Change*, *5*(5), 4.

Marsh, H. W. (1980). The influence of student, course, and instructor characteristics in evaluation of university teaching. *American Educational Research Journal*, *17*(1), 219-237.

Marsh, H. W. (1981). Prior subject interest, students' evaluations, and instructional effectiveness. *Multivariate Behavioral Research*, *16*, 83-104.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*, 77-95.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52*, 1187-1197.

McKeachie, W. J. (1997a). Student evaluations of teaching. *The professional evaluation of teaching* (American Council of Learned Societies Occasional Paper No. 33) [Online]. Available: http://www.acls.org/op33.htm

McKeachie, W. J. (1997b). Student ratings: The validity of use. *American Psychologist*, *52*, 1218-1225.

Murkison, G. (1991). The search for a smoking gun in student evaluations: Maybe there is one! *Proceedings of the annual meeting of the Institute for Management Sciences, Southeastern Chapter*, *27*, 1-4.

Pickett, J. R. (1987). Do grades influence student evaluations? An empirical evaluation. *Proceedings of the annual meeting of the Institute for Management Sciences, Southeastern Chapter*, *23*, 357-358.

Randall, C. H., Price, B. A., Tudor, L., & Stapleton, R. J. (1999). In search of a better mousetrap: Can a composite profile accurately evaluate teaching effectiveness? *Proceedings of the National Decision Science Institute annual meeting*, *30*, 266-268.

Rawls, J. (1999). *A theory of justice*. Cambridge, MA: Harvard University Press.

Stapleton, R.J. (1990). Academic entrepreneurship: Using the case method to simulate competitive business markets. *The Organizational Behavior Teaching Review*, *14* (4), 88-104.

Stapleton, R. J., & Stapleton, D. C. (1996). Randomly selecting students to lead case method discussions: Problems and pitfalls in performance appraisal. *Proceedings of the annual meeting of the Institute for Operations Research and the Management Sciences, Southeastern Chapter*, *32*, 117-119.

Stapleton, R. J., & Stapleton, D. C. (1998). Teaching business using the case method and transactional analysis: A constructivist approach. *Transactional Analysis Journal*, *28*, 157-167.

Tang, T.L.-P. (1997). Teaching effectiveness at a public institution of higher education: Factors related to the overall teaching effectiveness. *Public Personnel Management*, *26*, 379-387.