Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Contemporary Educational Psychology xxx (2004) xxx–xxx
Contemporary Educational Psychology
www.elsevier.com/locate/cedpsych

Grading leniency, grade discrepancy, and student ratings of instruction

Bryan W. Griffin*

Department of Curriculum, Foundations, and Reading, Georgia Southern University, P.O. Box 8144, Statesboro, GA 30460, USA

Abstract

The purpose of this study was to examine how grading leniency and grade discrepancy (the difference between expected grades and deserved grades) were associated with various dimensions of student ratings of instruction. A sample of 754 undergraduate college students completed a student ratings of instruction instrument and provided responses to a number of other questions on topics such as course difficulty and workload. A series of multilevel regression analyses were conducted and results showed that an instructor's grading leniency, as perceived by students, was positively associated with student ratings on all dimensions of instruction examined. This finding suggests that more lenient instructors tend to receive higher student ratings. The second finding shows that grade discrepancy was negatively associated with most dimensions of instruction. This supports the self-serving bias hypothesis under attribution theory (Gigliotti & Buchtel, 1990) in that students tended to punish instructors with lower ratings when expected grades were lower than students believed they deserved, yet little evidence of a pattern of rewards existed in student ratings when students expected grades higher than they deserved.

Keywords: Student ratings of instruction; Student evaluations of instruction; Grading leniency; Grade discrepancy; Self-serving bias; Attribution theory

* Fax: 1-912-681-5382.
E-mail address: bwgriffin@gasou.edu.

0361-476X/$ - see front matter © 2004 Published by Elsevier Inc.
doi:10.1016/j.cedpsych.2003.11.001

1. Introduction

Student ratings are widespread and a common tool for evaluating faculty. When asked, most faculty members approve of the use of student ratings of instruction for teaching improvement (Baxter, 1991; Griffin, 1999; Moses, 1986; Schmelkin, Spencer, & Gellman, 1997), but many are resistant to the use of student ratings for tenure, promotion, and merit decisions (Feldman, 1997; McKeachie, 1997a). What many educators believe is that student ratings are affected, or biased, by a number of factors unrelated to teaching performance (Marsh & Overall, 1979; Wilson, 1998), and one common concern is that grading standards employed by instructors could bias ratings. As Marsh and Roche (2000) have noted, the average correlation between expected grades and student ratings of instruction is around .20. Typically this relationship has been interpreted using one of three theoretical explanations (for reviews see Greenwald & Gillmore, 1997a; Marsh & Roche, 2000; Wachtel, 1998).

First, the positive correlation between expected grade and student ratings of instruction may be explained as indicating a valid measurement of student ratings since better instruction should result in more learning, better grades, and better ratings. Second, the association between expected grades and ratings of instruction could be spurious and produced by various student characteristics such as motivation. For example, more motivated students who have greater interest in the subject matter are likely to learn more, achieve more, and rate the instructor higher. Third, an association between expected grades and ratings could reflect some type of biasing effect. For example, one possible biasing effect is grading leniency. Under this hypothesis, instructors are rewarded with higher ratings for assigning higher grades as a result of lenient grading practices, or conversely penalized with lower ratings for assigning lower grades due to grading harshness. One important weakness of studies examining the grading leniency hypothesis is that few have incorporated measures of student perceptions of the instructor's grading leniency (Marsh, 1987; Marsh & Roche, 2000).

Olivares' (2001) was the only study found that incorporated a measure of grading leniency. Olivares measured grading leniency by asking students to compare their current instructor to others they have had and rate this instructor's grading from 1 "much easier/lenient grader" to 7 "much harder/strict grader." Olivares found zero-order correlations of −.42 between grading leniency and an overall rating of the instructor, and of −.45 between grading leniency and a composite rating of the instructor based on students' perceptions of the instructor's organization, communication, level of caring, and classroom atmosphere. Given the scoring system of the rating scale used for grading leniency, the negative correlations indicate that more lenient grading was associated with higher ratings of the instructors. Olivares also found that the association between grading leniency and student ratings of the instructor remained after controlling for pre-course interest, change in interest, expected grade for the course, and a measure of cognitive ability.

In addition to the grading leniency hypothesis, another possible biasing effect interpretation for the grades–ratings association can be found in the theories of attribution and retribution (Feldman, 1997). Attribution theory suggests that a student may react in one of two ways if that student receives a grade that differs from what was expected. If the grade is lower than expected, then the student is likely to activate a defensive mechanism commonly referred to as self-serving bias (Gigliotti & Buchtel, 1990). With self-serving bias, a student will attempt to protect his or her view of self and assign blame for the lower than expected performance to an external cause. The likely target will be the instructor, so the student will rate the instructor lower, thus a rating penalty effect will occur. If a student receives a grade that is higher than expected, then the student will assign credit to this performance to internal causes, such as his or her intelligence, ability, hard work, etc. Since the better than expected grade is seen as a result of the student's behavior or ability, ratings of the instructor are not likely to differ from ratings given by students who receive grades as expected; in essence, there is no rating reward effect. Further diminishing the possible rating reward effect is the situation identified by Miller and Ross (1975) in which individuals typically anticipate positive outcomes, so it is unlikely that many students will acknowledge higher than expected grades since high grades were expected anyway. In short, with attribution theory and self-serving bias, students are likely to penalize instructors for lower than expected grades, but there is unlikely to be any reward effect for the few students who might believe they are receiving a grade higher than expected. Retribution effect (Feldman, 1997) predicts simpler behavior on the part of students. If, for example, a student receives lower than expected grades, this individual will penalize the instructor, while a student who receives higher than expected grades will reward the instructor.

One difficulty with student ratings research using the self-serving bias and retribution effect explanations has been the method for determining the grade discrepancy—whether grades are higher or lower than what students expect. The most direct method for assessing grade discrepancy is usually found in grade manipulation experiments in which students are lead to anticipate one grade, but then receive a grade inconsistent with their expectations (e.g., Abrami, Dickens, Perry, & Leventhal, 1980; Tata, 1999; Worthington & Wong, 1979). Reviewers of these studies, however, have pointed to a number of potential flaws. One important flaw is that in classroom settings, often students do not know what their actual grade will be before they complete instructional rating forms, so the external validity of these studies is limited. For correlational studies of attribution and retribution effects, researchers often calculate grade discrepancy by considering pre-course grade point average (GPA) or pre-course expected grade, and then examining how the end-of-course expected grade or actual grade differs from the pre-course GPA or expected grade (e.g., Gigliotti & Buchtel, 1990; Granzin & Painter, 1973; Greenwald & Gillmore, 1997b; Palmer, Carliner, & Romer, 1978). A potential limitation of these designs is that students are very likely to reassess their expectations once they are exposed to the course and instructor, so pre-course grade expectation may provide an inaccurate grade discrepancy baseline. Similarly, the use of GPA for determining grade discrepancy could be misleading since performance, and expectation for performance, in a given course can be independent of performance in other courses. This does not mean that previous correlational studies are flawed or misleading, but alternative methods for assessing grade discrepancy may prove useful.

The purpose of this study is twofold. First, since only one study of the grading leniency hypothesis has incorporated a measure of leniency as perceived by students, it is important to understand better how scores from such a measure relate to student ratings, and to learn if the association between grading leniency and student ratings replicates across studies. Second, the calculation of grade discrepancy for assessing the self-serving bias and retribution effect hypotheses can be done in a manner that is perhaps more course appropriate than previously examined. Thus, the intent of this study is to examine the grading leniency explanation of student ratings by incorporating a measure of students' perceptions of leniency, and to test both self-serving bias and retribution effect hypotheses by incorporating a more course specific measure of grade discrepancy.

Method

Participants

A total of 754 undergraduate students enrolled in 39 education courses at a medium sized (14,000 students), Regional University in the southeastern United States participated in this study. The classes ranged in size from 6 to 34 students. Undergraduate education students at this institution are predominately White (71%) and female (80%). Most respondents (76%) reported grade point averages in the range of 2.5–3.5 on a 4.0 scale. Data were collected during the fall and spring semesters of the 1998–1999 academic year.

Instrument and variables

An instrument to assess student evaluations of instruction and course characteristics was developed drawing item and question wording from multiple sources (Abrami, d'Apollonia, & Rosenfield, 1997; Feldman, 1997; Marsh, 1987; Murray, 1997). To measure teaching effectiveness, 12 statements were used to assess multiple dimensions of instruction with ratings following a five-point scale. The 12 statements follow.

1. Overall, how would you rate this course?
2. Overall, how would you rate this instructor?
3. The instructor was dynamic and energetic in conducting the course.
4. The instructor presented the material in a clear and understandable manner.
5. Course materials were well prepared and organized.
6. Students were invited to share their ideas and knowledge.
7. The instructor made students feel welcome in seeking help/advice in or outside of class.
8. The content of this course is useful, worthwhile, or relevant to you.
9. Methods of evaluating student work were fair and appropriate.
10. The instructor seems to have a real interest in and concern for students.
11. The instructor gave students useful/helpful feedback on work.

12. The instructor is very knowledgeable in the subject of this course.

For the first 2 items, overall course and overall instructor, the scale ranged from 1 "Poor" to 5 "Excellent" and for the remaining 10 items the scale ranged from 1 "strongly disagree" to 5 "strongly agree."

The two predictors of interest in this study are grading leniency, which was assessed by students' responses to this statement, "This instructor is a lenient/easy grader" (1 "strongly disagree" to 5 "strongly agree"), and grade discrepancy, which was calculated as the difference between the grade a student expected ("What grade do you think the instructor will assign you in this course?") minus the grade a student believed they deserved in the course ("What grade do you think you deserve in this course?"). Both expected and deserved grades were assessed using a 12-point scale (i.e., $A+ = 13$, $A = 12$, $A- = 11$, etc. through $D- = 2$, $F = 1$). The difference between expected minus deserved grade can be interpreted as follows: a positive difference indicates the expected was higher than the deserved grade (e.g., expect an *A*− but deserve a *B*+), no difference shows expected and deserved are the same (e.g., expect a *B* and deserve a *B*), and a negative difference shows that expected grade is lower than deserved grade (e.g., expect *B*+ and deserve *A*−).

In addition to these measures, students also provided information concerning: (a) the instructor's reputation (1 "very bad" to 5 "very good," and 6 "didn't know about the instructor"), (b) course difficulty (1 "one of easiest" to 5 "one of most difficult"), (c) course workload (1 "very light" to 5 "very heavy"), (d) current GPA, and (e) pre-course motivation ("You had a strong desire to take this course," with responses ranging from 1 "strongly disagree" to 5 "strongly agree"). Class size and instructor's sex were also included in the analysis. Three categories of instructor reputation were developed for the analyses performed in this study: negative reputation, which included students who selected responses 1–3 ("instructor very bad" to "about average") for the instructor reputation item; positive reputation, which included students who choose responses 4 and 5 ("above average" to "instructor very good") for the instructor reputation item; and no information, which consisted of students who selected response 6 ("didn't know about the instructor") for the instructor reputation item.

From these three categories of instructor reputation, two dummy variables (Pedhazur, 1997) were created for the regression analyses performed below. The first, called positive reputation, was coded 1 if student responses corresponded with the positive reputation category, and 0 otherwise. The second dummy variable was labeled negative reputation and was coded 1 if student responses corresponded with the negative reputation category, otherwise a 0 was used. Of the 754 respondents, 176 (23.3%) were classified into the positive reputation group, 420 (55.7%) into the no information group, and 158 (21%) into the negative reputation group.

Evidence for construct validity for the scores obtained from this instrument and sample can be assessed by examining correlations among scores from the dimensions of instruction and various other course-related variables. Correlations and descriptive statistics for the student-level variables are presented in Table 1. For example, prior research has demonstrated a generally positive relationship between students' pre-course motivation and students' ratings of instruction (Marsh, 1987), and a

Table 1
Descriptive statistics and correlations among student-level variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	1.000																				
2	.789	1.000																			
3	.715	.646	1.000																		
4	.716	.676	.728	1.000																	
5	.654	.618	.700	.759	1.000																
6	.479	.391	.499	.492	.481	1.000															
7	.617	.505	.548	.556	.544	.652	1.000														
8	.556	.658	.569	.608	.543	.415	.498	1.000													
9	.604	.519	.543	.571	.570	.602	.637	.467	1.000												
10	.675	.557	.635	.628	.618	.644	.762	.511	.703	1.000											
11	.656	.577	.645	.664	.662	.604	.663	.520	.671	.737	1.000										
12	.544	.487	.605	.597	.627	.537	.504	.506	.571	.658	.649	1. 000									
13	.232	.167	.158	.190	.164	.238	.252	.061	.361	.276	.239	.145	1. 000								
14	−.005	−.051	.033	.007	.003	.049	.062	−.048	.050	.031	.016	−.008	.055	1.000							
15	−.235	−.217	−.147	−.201	−.153	−.171	−.249	−.132	−.310	−.242	−.205	−. 087	−.213	−.114	1.000						
16	.216	.202	.150	.176	.153	.109	.127	.174	.122	.134	.120	.117	.084	. 012	−.058	1.000					
17	−.356	−.304	−.213	−.240	−.199	−.256	−.294	−.163	−.369	−.319	−.284	−. 218	−.253	−.053	.229	−.284	1.000				
18	.131	.135	.157	.081	.133	.072	.045	.178	.027	.075	.099	.174	−.337	−. 101	.199	.012	.112	1.000			
19	.048	.057	.114	−.021	.083	.015	.023	.039	.021	.039	.089	.086	−.169	−. 048	.094	−.047	.067	.478	1.000		
20	.366	.496	.361	.381	.341	.203	.276	.486	.293	.304	.349	.254	.113	−. 034	−.073	.143	−.141	.098	.134	1.000	
21	.166	.164	.153	.166	.130	.153	.226	.089	.244	.189	.172	.085	.158	. 094	−.431	.051	−.174	−.275	−.110	.116	1.000
M	3.86	3.50	4.06	3.87	4.12	4.52	4.26	4.04	4.19	4.27	4.13	4.47	2.94	0. 03	0.29	0.23	0.21	3.25	3.47	3.21	10.54
SD	1.16	1.13	1.11	1.15	1.02	0.81	0.99	1.14	1.02	0.97	1.01	0.80	1.16	0. 17	0.46	0.42	0.41	0.90	0.94	1.10	1.77

Note. Variables include: 1, Overall Instructor Rating; 2, Overall Course Rating; 3, Dynamic/Energetic Rating; 4, Presented Clearly Rating; 5, Materials Organized Rating; 6, Students Invited to Share Ideas Rating; 7, Students Could Seek Help Rating; 8, Course Content Worthwhile Rating; 9, Fair Evaluations Rating; 10, Instructor Show Interest in Students Rating; 11, Feedback Helpful Rating; 12, Instructor Knowledgeable Rating; 13, Grading Leniency; 14, Positive Discrepancy (coded 1 if grade higher than deserved, 0 otherwise); 15, Negative Discrepancy (coded 1 if grade lower than deserved, 0 otherwise); 16, Positive Reputation Dummy (1 if student rated instructor as having positive reputation, 0 otherwise); 17, Negative Reputation Dummy (1 if student rated instructor as having negative reputation, 0 otherwise); 18, Course Difficulty; 19, Course Workload; 20, Pre-course Motivation; 21, Expected Grade.

All correlations larger than .071 in absolute value are statistically significant at the .05 level.

$n = 754$.

similar pattern emerges for these data. Additionally, the grade students expect for a course correlates positively with ratings for the course (Wachtel, 1998), and this pattern also can be observed with these ratings. Similar findings exist for course workload and course difficulty (Greenwald & Gillmore, 1997a, 1997b; Marsh & Roche, 2000).

Procedures

Students in 39 classes were administered the evaluation instrument during the last week of regular classes in the fall and spring semesters of the 1998–1999 academic year. Instructors were required to leave the classroom during evaluations. Students were told that evaluations would not be made available until after course grades had been assigned and would only be provided to instructors in aggregate form.

Results

Of the 754 students sampled, 67.8% ($n = 511$) believed that the grade they expected in the course was the grade they deserved, hence there was no difference between expected and deserved grade for these students. A total of 222 students (29.4%) expected a grade lower than they deserved and only 23 students (3.1%) expected a grade higher than they deserved. Of the two competing theories, self-serving bias and retribution effect, these data provide a better fit to the self-serving bias explanation since so few students surveyed thought they were to receive a grade higher than deserved. Miller and Ross (1975) predicted such behavior. It is also interesting to note that the majority of students expected no discrepancy at all, so it is likely that any grade discrepancy effect on student ratings of instruction may be small or limited to only a minority of students overall.

To statistically model student ratings, it was necessary to create dummy variables (Pedhazur, 1997) for grade discrepancy. The first, labeled positive discrepancy, was created to represent those students who believed their expected grade would be higher than deserved. The coding for this dummy was 1 for students expecting grades higher than deserved, and 0 for all other students. The second dummy variable, called negative discrepancy, was created to represent those students who believed their expected grade would be lower than their deserved grade, with coding of 1 for students expecting lower grades, and 0 for all others.

As the correlations in Table 1 show, grading leniency was positively correlated with each of the 12 instructional rating items. The correlations ranged from a low of .06 to a high of .36, with an average correlation of .21. The positive discrepancy dummy variable showed an inconsistent pattern of correlations, with both positive and negative correlations with the 12 ratings items, and with no correlation greater than .06 in absolute value. The negative discrepancy dummy demonstrated a consistently negative pattern of correlations with each of the 12 ratings items, with correlations ranging from −.08 to −.31. These correlations indicate that students with lower expected than deserved grades tended to rate the instructor and instruction lower on each of the 12 instructional rating items.

While the zero-order correlations are informative about the general nature of the relationship among these variables, it is important to determine whether these patterns of association remain once other predictors of student ratings are taken into account in a regression equation. To learn whether grading leniency and grade discrepancy are associated with student ratings of instruction, multilevel regression (Bryk & Raudenbush, 1992; Goldstein, 1995; Longford, 1993) was used in an effort to examine variation in student ratings both within and across classes. Several researchers of student ratings of instruction (e.g., Cranton & Smith, 1990; Feldman, 1998; Gigliotti & Buchtel, 1990) have noted that the level of analysis, either student- or class-level, at which student ratings are examined could influence the nature of statistical relationships revealed. For example, the analysis of class means rather than student-level data may obscure important variation in ratings that result from individual student differences within the classroom. Multilevel analysis allows one to combine both levels of analysis to provide a more complete model of student ratings.

Incorporated into the multilevel analyses that follow were several covariates previously identified as important predictors of student ratings of instruction. At the student level, these covariates include course difficulty, course workload, pre-course motivation, instructor reputation, and expected grade in the course. Research on student ratings has demonstrated course difficulty and course workload, often measured together, to correlate positively with ratings of instruction (Greenwald & Gillmore, 1997a, 1997b; Marsh, 1980; Marsh & Roche, 2000). Interest in the subject matter of the course before enrollment—pre-course motivation—has been linked to higher student ratings of instruction (Howard & Maxwell, 1980; Marsh, 1980; Prave & Baril, 1993). Barké, Tollefson, and Tracy (1983), Griffin (2001), and Ory (1980) found that instructor reputation was associated with various measures of teaching effectiveness. Finally, expected grade in the course, which typically correlates positively with ratings, has been the subject of much debate and research (Greenwald & Gillmore, 1997a; Marsh, 1987; Marsh & Roche, 1997, 2000; McKeachie, 1997b) and therefore was included in the analysis.

At the class level, class size and instructor sex were included. Research shows that class size correlates, albeit weakly, with ratings of instruction (Feldman, 1994). The sex of the instructor also appears to relate to student ratings. Feldman's (1998) reviews have shown that women tend to receive slightly higher ratings than men. However, Feldman (1998) also notes that a same-sex favorability in ratings exists; students of the same sex as their instructor may provide slightly higher ratings (Centra & Gaubatz, 2000). Since the majority of students in the classes examined in this study were women, it is likely that women instructors in this sample may have higher ratings.

Thus, the models examined were, with variables enclosed in parentheses, as follows:

Student-level

$$\begin{aligned}(\text{Student Rating of Instruction Item})_{ij} = {}& \beta_{0j} + \beta_1(\text{Grading Leniency})_{ij}\\ &+ \beta_2(\text{Positive Discrepancy})_{ij}\\ &+ \beta_3(\text{Negative Discrepancy})_{ij}\\ &+ \beta_4(\text{Positive Reputation})_{ij}\\ &+ \beta_5(\text{Negative Reputation})_{ij}\\ &+ \beta_6(\text{Course Difficulty})_{ij}\\ &+ \beta_7(\text{Course Workload})_{ij}\\ &+ \beta_8(\text{Pre-course motivation})_{ij}\\ &+ \beta_9(\text{Expected Grade})_{ij} + e_{ij}.\end{aligned}$$

At the class-level, mean ratings of the instructor were modeled with class size and instructor sex:

Class-level

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Instructor's Sex})_j + \gamma_{02}(\text{Class Size})_j + \mu_{0j}.$$

Combining the student- and class-level equations yields the following model of instructor rating:

Combined

$$\begin{aligned}(\text{Student Rating of Instruction Item})_{ij} = {}& \gamma_{00} + \beta_1(\text{Grading Leniency})_{ij}\\ &+ \beta_2(\text{Positive Discrepancy})_{ij}\\ &+ \beta_3(\text{Negative Discrepancy})_{ij}\\ &+ \beta_4(\text{Positive Reputation})_{ij}\\ &+ \beta_5(\text{Negative Reputation})_{ij}\\ &+ \beta_6(\text{Course Difficulty})_{ij}\\ &+ \beta_7(\text{Course Workload})_{ij}\\ &+ \beta_8(\text{Pre-course motivation})_{ij}\\ &+ \beta_9(\text{Expected Grade})_{ij}\\ &+ \gamma_{01}(\text{Instructor's Sex})_j\\ &+ \gamma_{02}(\text{Class Size})_j + e_{ij} + \mu_{0j}.\end{aligned}$$

This combined model was used to estimate the regression coefficients for each of the 12 rating items presented above. Multilevel regression results, using full information maximum likelihood to obtain estimates (Hox, 1995), are presented in Table 2.

Table 2
Multilevel regression results for student ratings of instruction

	Overall Instructor		Overall Course		Dynamic and Energetic		Presented Clearly		Materials Organized		Students Shared Ideas	
	B	*SE B*	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*
Fixed Portion of Model												
Student Level												
Grading Leniency	.12*	.03	.06*	.03	.11*	.03	.12*	.03	.10*	.03	.12*	.03
Grade Discrepancy												
Positive Discrepancy	−.14	.17	−.32	.17	.19	.18	.06	.18	.01	.18	.20	.15
Negative Discrepancy	−.24*	.07	−.23*	.07	−.10	.07	−.21*	.08	−.11	.08	−.11	.06
Instructor Reputation												
Positive Reputation	.21*	.08	.10	.07	.08	.08	.07	.08	.09	.08	.04	.07
Negative Reputation	−.39*	.10	−.32*	.09	−.19*	.10	−.08	.10	−.13	.10	−.28*	.08
Course Difficulty	.17*	.04	.13*	.04	.13*	.04	.13*	.05	.11*	.04	.15*	.04
Course Workload	.00	.04	.02	.04	.01	.04	−.07	.04	.03	.04	−.03	.04
Pre-course Motivation	.20*	.03	.32*	.03	.18*	.03	.20*	.03	.17*	.03	.08*	.03
Expected Grade	.08*	.02	.07*	.02	.07*	.02	.10*	.02	.06*	.02	.06*	.02
Intercept	2.05*	.46	1.80*	.42	2.30*	.46	2.40*	.47	2.65*	.42	2.97*	.33
Class Level												
Class Size	−.01	.01	−.02	.01	−.01	.01	−.02	.01	−.01	.01	.00	.01
Instructor's Sex	−.54*	.21	−.41*	.18	−.46*	.20	−.41*	.20	−.36*	.16	−.10	.11
Random Portion of Model												
Class-level variance	.35*		.27*		.35*		.33*		.20*		.08*	
Student-level variance	.62*		.57*		.64*		.68*		.64*		.48*	
R^2 (total variance modeled)	.32		.36		.22		.24		.20		.17	

Table 2 (*continued*)

	Students Could Seek Help		Course Content Worthwhile		Fair Evaluation of Students		Interest in Students		Feedback Helpful		Instructor Knowledgeable	
	B	*SE B*	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*
Fixed Portion of Model												
Student Level												
Grading Leniency	.13*	.03	.03	.03	.19*	.03	.13*	.03	.14*	.03	.08*	.03
Grade Discrepancy												
Positive Discrepancy	.26	.18	−.12	.19	.12	.17	.10	.17	.06	.17	−.01	.15
Negative Discrepancy	−.25*	.08	−.19*	.08	−.31*	.07	−.23*	.07	−.17*	.07	−.01	.06
Instructor Reputation												
Positive Reputation	.05	.08	.05	.08	−.02	.07	.04	.07	.02	.08	.00	.07
Negative Reputation	−.37*	.09	−.27*	.10	−.49*	.09	−.36*	.09	−.30*	.09	−.32*	.08
Course Difficulty	.15*	.04	.18*	.05	.13*	.04	.14*	.04	.11*	.04	.16*	.04
Course Workload	−.01	.04	−.01	.04	.02	.04	.02	.04	.07	.04	.01	.04
Pre-course Motivation	.15*	.03	.37*	.03	.16*	.03	.16*	.03	.20*	.03	.09*	.03
Expected Grade	.07*	.02	.04	.02	.08*	.02	.06*	.02	.08*	.02	.04*	.02
Intercept	2.54*	.38	2.43*	.44	2.05*	.38	2.69*	.38	2.10*	.40	3.39*	.33
Class Level												
Class Size	−.01	.01	−.01	.01	−.01	.01	−.01	.01	−.01	.01	−.02	.01
Instructor's Sex	−.37*	.12	−.43*	.17	−.15	.14	−.29*	.14	−.30*	.15	−.20	.11
Random Portion of Model												
Class-level variance	.09*		.23*		.13*		.14*		.16*		.08*	
Student-level variance	.65*		.71*		.57*		.56*		.62*		.46*	
R^2 (total variance modeled)	.26		.31		.32		.27		.25		.17	

Note. Positive Discrepancy coded 1 if expected grade is higher than believed deserved, 0 otherwise; Negative Discrepancy coded 1 if expected grade lower than believed deserved, 0 otherwise; Positive Reputation dummy coded 1 if student rated instructor as having positive reputation, 0 otherwise; and Negative Reputation dummy coded 1 if student rated instructor as having negative reputation, 0 otherwise. R^2 is calculated in the normal manner (Pedhazur, 1997), but model variance is calculated by summing both the between and within class variances (Snijders & Bosker, 1999).

$n = 754$ students in 39 courses.

$^* p < .05$.

The regression results in Table 2 indicate that grading leniency was statistically and positively related to 11 of the 12 rating items. The weakest relationship ($b = .03$) was with the course content item, and this was the only partial coefficient for grading leniency that was not statistically significant. The strongest relationship ($b = .19$) was with the fair evaluation of students item. The latter coefficient may be interpreted as showing that the more lenient the instructor's grading, the more fair and appropriate was judged the instructor's evaluations of students' work. The average partial regression coefficient for the 12 items was .11. To put these estimates into perspective, consider the situation of examining the single overall instructor rating item for which the grading leniency regression estimate is $b = .12$. Assuming that all other factors are held constant, two instructors who differ only on perceived grading leniency by one standard deviation ($SD = 1.16$, see Table 1) could expect an average mean difference of $1.16 \times .12 = .14$ points on their overall instructor rating item. On the extremes, one instructor judged the least lenient (rating = 1) and another judged most lenient (rating = 5) would differ by $(5–1) \times .12 = .48$ points on their average overall instructor rating; for example, say 4.48 vs. 4.00 on a scale of 1–5.

The relationship between grade discrepancy and student ratings was more complex than that found with grading leniency. The positive discrepancy dummy variable was positively related to 8 of the 12 ratings items, and negatively related to the remaining 4 ratings items. In no cases were the coefficient estimates for this dummy variable statistically significant, and in all cases the standard errors for the coefficients were relatively large, thus indicating great variability in the estimates. Given the small sample size of students who thought their expected grade was higher than their deserved grade ($n = 23$), such unreliable estimates should be expected. The regression estimates obtained for the positive discrepancy dummy show a weak and inconsistent pattern of rating behavior for this group of students.

Unlike the positive discrepancy dummy, the dummy variable negative discrepancy demonstrated a consistent and negative pattern of rating behavior for students expecting grades lower than they perceive they deserved. The negative discrepancy dummy was found to be negatively associated with student ratings in all cases, and was statistically significant for 8 of the 12 ratings items. Since negative grade discrepancy is a dummy variable, the regression coefficient may be interpreted as the mean difference in student ratings between those students who expect a grade lower than they deserve and everyone else. The largest difference ($b = -.31$) was for the fair evaluation of students item, and the smallest difference ($b = -.01$) was found for the instructor knowledgeable item. Drawing on the example above using the overall instructor rating item, consider two instructors who differ only in the expectations held by their students regarding their expected and deserved grades. The overall instructor rating for the instructor with students who believe their expected grades will be lower than they deserve will be −.24 points lower than the instructor whose students do not anticipate any difference between their expected and deserved grades, e.g., 4.00 vs. 4.24.

For the other variables included in the models, results mirrored findings from previous studies. The strongest predictor of ratings was pre-course motivation. The negative instructor reputation dummy variable was negatively related to each rating item except for two. Course difficulty was consistently, and positively, related to all rating items. The more difficult the course, as judged by students, the more positive were student ratings. Course workload was not statistically related to any of the rating items. Expected grade was also positively and statistically related to 11 of the 12 rating items. The partial regression coefficients for expected grade ranged from a low of .04 to a high of .10.

Discussion

Recall the three possible interpretations of the positive relationship between expected grade and student ratings of instruction: (a) valid teaching/learning association, (b) spurious association, and (c) biasing effect. Two ways of expressing the biasing effect were examined in this paper, grading leniency and grade discrepancy. Grading leniency was positively, and linearly, associated with 11 of the 12 rating items. The positive relationship means that students tended to rate higher those instructors judged to be more lenient graders, and, conversely, instructors with harsher grading practices tend to receive lower ratings. This finding replicates that reported by Olivares (2001) who also found that instructors with more lenient grading practices tended to have higher student ratings. On the basis of results from this study and Olivares' study, it appears that students rate instructors who are lenient graders higher than instructors who are less lenient with their grading.

Also examined was the relationship between student ratings and grade discrepancy, which is defined in this study as the difference between students' expected grade and perceived deserved grade. Two theoretical explanations for such an effect were listed, self-serving bias and retribution effect. As noted, self-serving bias suggests that students will penalize instructors for lower than deserved grades, but will not reward instructors for higher than deserved grades. Retribution effect holds that students will reward instructors for higher than deserved grades, and penalize instructors for lower than deserved grades. The data examined here provide a better fit to the self-serving bias hypothesis. Only about 3% of the students sampled expected grades higher than they deserved, and about 29% expected grades lower than they deserved. There was little evidence that those who expected higher than deserved grades rewarded instructors with higher ratings when compared to ratings made by other students in the sample. None of the regression estimates for this group of students was statistically different from zero. There is, however, evidence of a penalty effect; students who expected grades lower than they deserved consistently provided ratings that were lower than other students. The differences, adjusted for the modeled covariates, ranged from low of −.01 to a high of −.31, with the overall average of −.18. This penalty effect is also consistent with findings of a grading harshness effect (Marsh & Roche, 2000; Worthington & Wong, 1979) in which students rate lower instructors perceived to grade harshly. Note, however, that Marsh and Roche (2000) point out that the self-serving bias may not be a bias under certain conditions for student ratings of instruction. Perhaps, for example, if a grade discrepancy is due to factors unrelated to instruction or the instructor, then students may not provide lower ratings. Unfortunately, the reason for a grade discrepancy was not assessed this study, so it is impossible to know further what students were thinking when they identified a grade discrepancy.

In summary, these results suggest two things. First, there may be a grading leniency effect in student ratings, but so far only this study and Olivares' (2001) study have apparently examined directly students' perceptions of grading leniency. Replication studies are needed to further evaluate this finding. Second, in addition to a possible grading leniency effect, there appears to be an association between a negative grade discrepancy and student ratings. This finding supports the self-serving bias hypothesis in that students appear to penalize instructors when grades are lower than expected, but do not reward instructors when grades are higher than expected. Since grading leniency and grade discrepancy, both possible parts of the biasing effect interpretation, were statistically controlled in the multilevel regression models, the partial regression coefficients for expected grade may represent a more pure examination of the: (a) valid teaching/learning association and (b) spurious association hypotheses. Several factors that could lead to the spurious association effect were included in the regression models, such as precourse motivation, course difficulty and workload. It is possible, though, that other factors could contribute to the observed relationship between expected grade and ratings found in this and other studies. More careful examinations taking into account various motivational factors such as intrinsic and extrinsic motivation, personal control, and autonomy may prove useful in further elimination of the spurious effects hypothesis. However, since at least part of the spurious association and biasing effects hypotheses have been controlled in this study, that means the relationships between expected grades and student ratings of instruction found in the current study probably can be explained, at least in part, by the valid teaching/learning hypothesis. Thus, the results provided here suggest that student ratings of instruction are probably a function of both valid teaching and learning and some biasing due to grading leniency and grade discrepancy.

References

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (1997). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 321–367). New York: Agathon.

Abrami, P. C., Dickens, W. J., Perry, R. P., & Leventhal, L. (1980). Do teacher standards for assigning grades affect student evaluations of instruction? *Journal of Educational Psychology, 72*, 107–118.

Barké, C. R., Tollefson, N., & Tracy, D. B. (1983). Relationship between course entry attitudes and end-of-course ratings. *Journal of Educational Psychology, 75*, 75–85.

Baxter, E. P. (1991). The TEVAL experience, 1983–88: The impact of a student evaluation of teaching scheme on university teachers. *Studies in Higher Education, 16*, 151–179.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 70*, 17–33.

Cranton, P., & Smith, R. A. (1990). Reconsidering the unit of analysis: A model of student ratings of instruction. *Journal of Educational Psychology, 82*, 207–212.

Feldman, K. A. (1994). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education, 21*, 45–116.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368–395). New York: Agathon.

Feldman, K. A. (1998). Reflections on the study of effective college teaching and student ratings: One continuing question and two unresolved issues. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 35–74). New York: Agathon.

Gigliotti, R. J., & Buchtel, F. S. (1990). Attributional bias and course evaluations. *Journal of Educational Psychology, 82*, 341–351.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Granzin, K. L., & Painter, J. J. (1973). A new explanation for students' course evaluation tendencies. *American Educational Research Journal, 10*, 115–124.

Greenwald, A. G., & Gillmore, G. M. (1997a). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*, 1209–1217.

Greenwald, A. G., & Gillmore, G. M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89*, 743–751.

Griffin, B. W. (1999). *Results of the faculty survey on student ratings of instruction: Preliminary report*. Statesboro, GA, USA: Georgia Southern University, Student Ratings Committee.

Griffin, B. W. (2001). Instructor reputation and student ratings of instruction. *Contemporary Educational Psychology, 26*, 534–552.

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.

Hox, J. J. (1995). *Applied multilevel analysis*. Amsterdam: TT-Publikaties. Available on-line (March 6, 2000): <http://www.ioe.ac.uk/multilevel/workpap.html>.

Longford, N. T. (1993). *Random coefficient models*. Oxford, UK: Oxford University Press.

Marsh, H. W. (1980). The influence of student, course, and instructor characteristics on evaluations of university teaching. *American Educational Research Journal, 17*, 219–237.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253–388.

Marsh, H. W., & Overall, J. U. (1979). *Validity of students' evaluations of teaching: A comparison with instructor self-evaluations by teaching assistants, undergraduate faculty, and graduate faculty*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document No. ED 177 205).

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.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92*, 202–228.

McKeachie, W. J. (1997a). Good teaching makes a difference—and we know what it is. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 396–408). New York: Agathon.

McKeachie, W. J. (1997b). Student ratings: The validity of use. *American Psychologist, 52*, 1219–1225.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin, 82*, 213–225.

Moses, I. (1986). Student evaluation of teaching in an Australian university staff perceptions and reactions. *Assessment & Evaluation in Higher Education, 11*, 117–129.

Murray, H. G. (1997). Effective teaching behaviors in the college classroom. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 171–204). New York: Agathon.

Olivares, O. J. (2001). Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology, 26*, 382–399.

Ory, J. C. (1980). The influence of students' affective entry on instructor and course evaluations. *The Review of Higher Education, 4*, 13–24.

Palmer, J., Carliner, G., & Romer, T. (1978). Leniency, learning, and evaluations. *Journal of Educational Psychology, 70*, 855–863.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Brace, New York: Harcourt.

Prave, R. S., & Baril, G. L. (1993). Instructor ratings: Controlling for bias from initial student interest. *Journal of Education for Business, 68*, 362–366.

Schmelkin, L. P., Spencer, K. J., & Gellman, E. S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education, 38*, 575–592.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Tata, J. (1999). Grade distributions, grading procedures, and students' evaluations of instructors: A justice perspective. *Journal of Psychology Interdisciplinary & Applied, 133*, 263–271.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*, 191–212.

Wilson, R. (1998). New research casts doubt on value of student evaluations of professors. *The Chronicle of Higher Education*, A12–A14.

Worthington, A. G., & Wong, P. T. P. (1979). Effects of earned and assigned grades on student evaluations of an instructor. *Journal of Educational Psychology, 71*, 764–775.