Research in Higher Education, Vol. 44, No. 5, October 2003 (© 2003)

WILL TEACHERS RECEIVE HIGHER STUDENT EVALUATIONS BY GIVING HIGHER GRADES AND LESS COURSE WORK?

John A. Centra*

This study investigated whether mean expected grades and the level of difficult/workload in courses, as reported by students, unduly influence student ratings instruction. Over 50,000 college courses whose teachers used the Student Instructional Report II were analyzed. In addition to the two primary independent variables, the regression analyses included 8 subject area groupings and controlled for such factors as class size, teaching method, and student perceived learning outcomes in the course. Learning outcomes had a large positive effect on student evaluations of instructions, as it should. After controlling for learning outcomes, expected grades generally did not affect student evaluations. In fact, contrary to what some faculty think, courses in natural sciences with expected grades of A were rated lower, not higher. Courses were rated lower when they were rated as either difficult or too elementary. Courses rated at the "just right" level received the highest evaluations.

KEY WORDS: student evaluations; college course ratings; expected grades; course difficulty.

INTRODUCTION

Some college teachers believe a sure way to win student approval is to give high grades and less course work. They further believe that this will translate into higher student evaluations of their teaching, a kind of quid pro quo. When faculty members at a major research university were asked what would most likely bias students evaluations, 72% said course difficulty, 68% reported grading leniency, and 60% reported course workload (Marsh, 1987). Are these faculty members in fact correct? Given the increased emphasis on using student evaluations of teaching in tenure and promotion decisions at many colleges, a teacher's temptation to manipulate grades or course workload is a possibility.

No method of evaluating college teaching has been researched more than student evaluations, with well over 2,000 studies referenced in the ERIC system.

*Address correspondence to: John A. Centra, 7811 Clearwater Circle, Manlius, NY 13104. E-mail: Jacentra@syr.edu

0361-0365/03/1000-0495/0

The preponderance of these study results has been positive, concluding that the evaluations are: (a) reliable and stable; (b) valid when compared with student learning and other indicators of effective teaching; (c) multidimensional in terms of what they assess; (d) useful in improving teaching; and (e) only minimally affected by various course, teacher, or student characteristics that could bias results (d'Apollonia and Abrami, 1997; Cashin, 1988; Centra 1993; Marsh, 1987; McKeachie, 1979). Spurred by the strong opinions held by many faculty members, several recent studies have addressed the effects of grading leniency and course workload on student evaluations. The results have been somewhat contradictory, perhaps because of limited sample sizes and the inability to include key variables such as the subject area of the course. The present study uses a large and diverse sample of college courses to investigate whether grades, more exactly the final grades students expect to receive at the completion of a course, have an undue influence on their ratings of teaching. In addition, the possible influences of the workload, difficulty level, and pace of the course on students' ratings, together with the subject area of the course and many other factors, will be studied.

Final grades in a course are typically not known to students at the time they complete a student evaluation form and thus should not be expected to bias their evaluations. Forms are generally filled out during the last few weeks of a course when students are aware of only the grade they think they will get based on their perceived performance to date. Therefore, studies of a possible grading-leniency bias on ratings can appropriately use expected grades. Not surprisingly, students' expected grades tend to show the same associations with other variables, as do the actual final grades (Franklin, Theall, and Ludlow, 1991).

Expected grades are correlated moderately with student evaluations: In a study of 9,194 class-average Student Instructional Report (SRI) responses from a variety of colleges and courses, expected grades and global ratings of teacher effectiveness correlated .20 (Centra and Creech, 1976). A more recent review of several studies by Feldman (1997) reported that correlations ranged between .10 and .30. Thus, in most studies and with most rating forms, whether rating the instructor or the course, the correlation has averaged close to .20. This moderate but significant relationship between expected grades and ratings has several possible explanations other than the quid pro quo one of grading leniency causing higher evaluation ratings. Foremost is the validity explanation: when students receive high grades in a course, it is a reflection of how well they have learned; they should therefore evaluate the course or teacher highly. This validity explanation will be the basis of the analysis used in this study, in which student perceived learning is controlled statistically. A second possible explanation for the expected grades/ratings correlation is based on students' academic motivation or their prior interest in the subject. Courses that attract strongly motivated or interested students should have higher grades because students

work harder and learn more. Those same courses should get higher ratings because motivated students appreciate the course and the instruction they have received (Howard and Maxwell, 1980; Marsh, 1987). A final explanation relies on attributional principles, whereupon people tend to accept credit for desired outcomes while denying responsibility for undesired outcomes (Greenwald, 1980). Thus, students would attribute their high grades to their hard work and intelligence, and low grades (and the ratings students give) would be attributed to poor instruction.

Experimental field studies investigating a grading-leniency bias on ratings used a design in which students were given false course grades (Holmes, 1972; Powell, 1977; Vasta and Sarmiento, 1979; Worthington and Wong, 1979). While these studies found some evidence of a grading-leniency effect, they contained weaknesses that make their conclusions questionable. As Marsh and Roche (1997) pointed out, the deception used by the researchers was not only ethically dubious but also violated students' reasonable grade expectations. Because actual and expected grades are typically correlated, giving students grades not related to their performance, and different from students in the class performing at the same level, was both unreasonable and offensive to students. In fact the negative effects of giving students undeserved higher grades was demonstrated in a study by Abrami, Dickens, Perry, and Leventhal (1980). Finally, experimenter bias (the researchers themselves usually taught the classes) and results that were statistically weak make any conclusions of grading-leniency bias unwarranted.

In summary, neither the field experimental studies nor the correlational data (i.e., the generally moderate correlations) provide convincing evidence for the conclusion that student ratings of courses were influenced by the grades they received from instructors. The previous studies, however, did not take into account important factors such as the subject field. It is well established that both grades and ratings vary by subject fields, with the humanities in particular rated higher and giving higher grades than the natural sciences (Cashin, 1990; Centra, 1993; Feldman, 1978). Therefore, although the overall correlation between ratings and grades may average only .20, correlations within subject fields may be much higher. This study investigates possible differences by subject fields along with other variables.

COURSE DIFFICULTY/WORKLOAD

Course difficulty/workload has frequently been measured by a combination of student ratings of the level of difficulty, workload, and pace of the course. In some instances, the number of hours students said they spent on the course outside of class was also included in a workload factor. Gillmore and Greenwald (1994) included hours per week along with three other items to imply workload: the students' reported challenge of the course, their effort, and their involvement.

This definition of workload seems to focus more on the individual student than the course-related set of items that assess difficulty, workload, and pace that have been used in other studies. It may in fact be their particular definition that resulted in workload being positively related to ratings as reported by Greenwald and Gillmore (1997). One other study with a large and diverse database demonstrated clearly that a student effort/involvement/challenge factor was highly correlated with evaluations of instruction (Centra and Gaubatz, 2000b).

Marsh and Roche's (2000) extensive study of grading leniency and workload effects on student evaluations used part of the usual workload definition (course difficulty, workload, and pace), but also included hours per week spent studying outside of class. As Greenwald and Gillmore (1997) found, higher workload was related to higher student evaluations (overall teacher $r = .19$, overall course $r = .25$). Thus teachers received higher ratings when they gave more work. A contrary finding, whereupon a lower workload was related to higher students' evaluations, and which some teachers believe happens, was reported by Franklin et al. (1991), although the effect size was small. The difference among these findings is likely due to the different definitions of workload. Franklin and associates did not include hours student reported studying outside of class, whereas the previous two studies did. Hours spent outside of class on coursework can be further refined by dividing it into good hours (deemed to be valuable by students) and bad hours (total hours minus good hours), as pointed out by both Gillmore and Greenwald (1994) and Franklin and Theall (1996). The importance of this distinction was underscored by Marsh (2001), who found that good hours were related to student evaluations of teachers and to their perceptions of learning; bad hours were negatively related to these same factors.

BIAS: WHAT IS IT?

One definition of bias that has been used in other studies of student evaluations of teaching (Centra and Gaubatz, 2000a) is as follows: Bias exists when a student, teacher, or course characteristic affects the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching, such as increased student learning. Class size, teacher experience, and teacher gender are examples of characteristics that correlated with student evaluations but are not necessarily biasing effects (Centra and Creech, 1976). Small classes with fewer than 15 students get higher evaluations than do larger classes, but if students learn more in smaller classes because they allow for more personal attention, then class size is not truly biasing the evaluations. Likewise, teachers in their first year of teaching generally receive lower evaluations than more experienced teachers, but because students may learn less from first-year teachers, the evaluations are not truly biased against these teachers. In a study of possible bias due to the gender of teachers and students, only small differences were found in

evaluations, and because these were related to self-reported student learning, bias did not exist (Centra and Gaubatz, 2000a).

Applying this definition of bias to the present study, it is important not only to investigate the relationships of expected grades and difficulty/workload to course evaluations but also to a measure of student learning in the course. While this study will not have available an objective measure of learning, such as final examination results, it will have a student self-reported learning measure. Several writers have supported students' self-reports of learning as an alternative to objective test results for validating student evaluations because they can tap a broader array of outcomes and attitudes, such as subject matter attitudes, motivational outcomes, and critical thinking (Dowell and Neal, 1982; Feldman, 1989; Koon and Murray, 1996). Moreover, studies have shown that self-reports of learning are reasonably correlated with the actual learning measures with which they overlap (Baird, 1976; Pike, 1995).

Thus in investigating the possible effects of grading leniency and course workload on student evaluations, this study controls for student self-reported learning outcomes. Unlike previous studies, several of which are contradictory, the analyses take into account many other possible influences on student evaluations: subject field of the course, class size, class level, course requirement, institutional type, teaching method, and student effort and involvement in the course. Each of these variables has been shown to be related to student evaluations, although in some instances only modestly (Centra, 1993, Marsh, 1987).

METHOD

The Student Instructional Report II (SIR II) was used in this study to measure student evaluations of instruction and other key variables. SIR II is a new version of the SIR, which was first made available to colleges by the Educational Testing Service in the early 1970s (Centra, 1972). The SIR II has some of the same instructional evaluation scales as the earlier version but has a new response format for students as well as new sets of questions to reflect more recent emphasis in college teaching. Its development and psychometric properties, including reliability and validity information, are described in Centra (1998).

The SIR II provided the primary independent variables in this study: students' self-reported expected grades and their evaluations of the difficulty/workload level of their courses. Grades were estimated by students on a 7-point scale, with 7 = A and 1 = below C (reversed from the questionnaire). The difficulty/workload for each course was an average of student responses to the three items listed in Table 1, with 5 = very elementary, much lighter, or very light (also a reversal from the questionnaire).

The wording of the items directs students to respond according to their own "preparation and ability" for the course difficulty question, or "in relation to

TABLE 1. Percentage of Each Response for the Course Difficulty, Workload, and Pace Items[a]

For my preparation and ability, level of difficulty of this course was:

Very Difficult	Somewhat Difficult	About Right	Somewhat Elementary	Very Elementary
8	31	54	5	1

The workload for this course in relation to other courses of equal credit was:

Much Heavier	Heavier	About the Same	Lighter	Much lighter
7	23	56	10	2

For me, the pace at which the instructor covered the material during the term was:

Very Fast	Somewhat Fast	Just About Right	Somewhat Slow	Very Slow
5	20	69	4	1

[a]Responses do not add to 100% because of omits.

other courses" for course workload. Also, unlike most other rating instruments, each of the five points on the scale is described, the midpoint of three being the most desirable response ("about right," "about the same"). This suggests a curvilinear relationship between the difficulty/workload level of courses and evaluations of instruction. A scatterplot of values did in fact show an inverted U curvilinear relationship, indicating that a quadratic as well as a linear application of the difficulty/workload variable was advisable. Moreover, a factor analysis of the three items resulted in a single factor, with each item having a factor loading of .86 or higher. Because the course difficulty item was the most dominant of the three (factor loading of .92), the appropriate description of this variable is course difficulty/workload. For the sake of brevity, this will at times be referred to simply as course difficulty in this study. As Table 1 indicates, the majority of classes were rated at the midpoint ("about right"). Fewer courses were rated as elementary, lighter (workload) or slower (pace) than were rated difficult, heavier, or faster.

Marsh and Roche (2000) defined workload with the same three items but with only the extremes of the response continuum described. They also included a question on the amount of study time students reported spending out of class.

As discussed earlier, Greenwald and Gillmore (1997) defined workload as course challenge plus time spent out of class. Both of these definitions differ from the course difficulty/workload variable in this study.

The dependent variables in this study were the instructional evaluation scales of SIR II and the overall evaluation of instruction item from the questionnaire (item 40). The SIR II scales had been validated through factor analysis and have excellent coefficient α and test-retest reliabilities (Centra, 1998). The Course Organization and Planning Scale (Scale A, 5 items) included the instructor's explanation of course requirements, use of class time, and emphasis of important points. The Communication Scale (Scale B, 5 items) included the instructor's ability to make clear presentations, use challenging questions or problems, and to be enthusiastic about the course material. Within the Faculty/Student Interaction Scale (Scale C, 5 items) were such items as the instructor's responsiveness to students, concern for student progress, and availability for help. Items in the Assignments, Exams, and Grading Scale (Scale D, 6 items) included the information given to students about how they would be graded, the clarity of exam questions, the instructors' comments on assignments and exams, and the helpfulness of assignments in understanding course material. Unique to the SIR II is that students responded to the items in these four scales and the overall evaluation item as each contributing to their learning. In short, the emphasis of the form is in tying practices to learning and in making students aware and responsive to that connection. And while most other forms' global or overall evaluation items ask students to rate the teacher or the course, the SIR II global item asks students to rate the quality of instruction as it contributed to their learning, using a linear 5-point effectiveness scale.

Student learning was assessed more directly with the Course Outcomes Scale (Scale F) by including the students' ratings of progress toward course objectives, increase in learning, increase in interest in the subject matter, the extent the course helped students to think independently about the subject, and the extent the course actively involved the student in what they learned.

A number of variables were also used as independent or control variables in addition to the primary ones (expected grades and course difficulty/workload). From the SIR II, student ratings of their Effort and Involvement (Scale G) formed a scale of three items, including amount studied and effort in the course, preparation for each class, and challenge of the course (all rated in relation to other courses on a 5-point scale). The Effort and Involvement Scale is similar to the Greenwald and Gillmore (1997) workload scale with its inclusion of challenge to students. It also correlated with the Course Outcomes Scale and to the Overall Evaluation of the course (Centra and Gaubatz, 2000b). It therefore made sense to control for this variable in the analyses. Other variables that entered the analyses as control variables and the codes used follow. Each instructor

provided the information on the "Instructor's Cover Sheet," with the exception of the Course Outcomes Scale:

- Institutional Type (0 = 2 year, 1 = 4 year or more)
- Class Size (0 = 16 or larger, 1 = 6–15)
- Class Level (0 = freshmen/sophomore, 1 = junior/senior)
- College Required Course vs. Student Choice (0 = college required, 1 = major/minor course or elective)
- Teaching Method of the Course (two variables) (1 = lecture/discussion, 0 = other; 1 = discussion or lab, 0 = other)
- Subject Area of the Course Grouped into Eight Categories (described in Table 4)
- The Course Outcomes Scale (5 = high, 1 = low)

SAMPLE

The sample for this study included a total of approximately 55,000 classes in which the SIR II had been administered from 1995 to 1999. Depending on the number of valid responses for each variable, analyses were based on between 46,182 to 55,549 classes. As Table 2 indicates, approximately 32% of these classes were in 2-year colleges and 68% were in 4-year colleges; 63% of the classes were in the students' major, minor, or an elective (37% were college required general education courses); 68% of the courses were at the junior or senior level and 32% were at the freshman or sophomore level; 68% of the courses had 16 students or more; the average expected grade across all classes was midway between a B and B+; the lecture teaching method was dominant, and the Overall Evaluation (item 40) as well as scales A through G were over 4.00, well above the numeric midpoint of 3.00. Course Outcomes, and Student Effort and Involvement were lowest at 3.71 and 3.69.

ANALYSIS

The class was the unit of analysis, rather than the individual student. Thus class average expected grade and class averages for all other variables were analyzed, resulting in a more reliable estimate of each variable and minimizing individual student variations. Hereafter, when expected grades and other variables are mentioned, it should be understood that they are all class average (mean) values. Stepwise multiple regression was the primary analysis used in this study. The dependent variables (Scales A, B, C, and D) plus the Overall Evaluation of the courses (item 40), were regressed on the 10 independent variables. Course Difficulty/Workload entered the analysis as both a linear and quadratic variable because of the finding of a single factor for the three items and

TABLE 2. Means and Standard Deviations of Variables ($N = 46{,}687–55{,}155$)[a]

	Mean	Standard Deviation
Scale A: Course org. and Planning	4.27	.46
Scale B: Communication	4.31	.44
Scale C: Fac./Student Interaction	4.31	.48
Scale D: Assignments, Exams, and Grading	4.09	.45
Scale F: Course Outcomes	3.71	.50
Scale G: Student Effort, Involvement	3.69	.41
Scale H: Course Diff., Work, Pace[b]	2.70	.35
Item 40: Overall Evaluation	4.02	.51
Item 41: College Required vs. Choice[c]	.63	.32
Class Level[d]	.32	.46
Class Size[e]	.32	.47
Institutional Type[f]	.68	.47
Expected Grade[g]	4.53	1.25
Teaching Method: Lecture/Discussion[h]	.59	.49
Teaching Method: Discussion/Lab[i]	.32	.47

[a]Scales A through F and Item 40, Overall Evaluation, were dependent variables; all others were independent variables, 5 = high, 1 = low.
[b]Mean of three items with 1 = difficult, fast, 3 = About right, 5 = elementary, slow.
[c]63% of classes in students' major, minor, or as electives.
[d]32% of classes at freshman/sophomore level; 68% Jr./Sr.
[e]1 = 6–15, 0 = 16 or larger.
[f]68% were 4-year colleges/universities; 32% were 2-year colleges.
[g]1 = below C, 7 = A, 4 = B, 5 = B+.
[h]59% classified by instructors as primarily lecture/discussion classes.
[i]32% classified by instructors as primarily discussions, labs, or labs with lectures; 9% classified by instructions as primarily lecture.

the evidence of some curvilinearity in the responses. Expected grade was also entered as both a linear and quadratic function due to a small degree of curvilinearity in the grade distributions. The eight subject area groupings (Table 4) were the third primary independent variable, requiring seven dummy indicators to represent them (the eighth, Health, did not require an indicator). Entering the regression first as control variables were the other independent variables: Student Effort and Involvement (Scale G), College Required Course vs. Student Choice, Class Size, Class Level, Institutional Type, Teaching by Lecture, Teaching by Discussion or Laboratories, and Course Outcomes. Due to the large N in this study, multicollinearity was not a problem, although not surprisingly some variables were highly correlated (Neter, Kutner, Nachtsheim, and Wasserman, 1996).

RESULTS

As the correlation matrix in Table 3 indicates, several of the scales of SIR II, together with the overall evaluation item (item 40), are highly intercorrelated. Because these are correlations of mean values, they are much higher than individual score correlations. This was true with previous analyses of SIR data as well, but a factor analysis revealed separate and distinct factors that provided useful information about instructional effectiveness (Centra, 1988). Of particular interest for this study are the correlations of student evaluations of instruction with expected grade and course difficulty/workload. Expected grade correlated only .11 with Overall Evaluation (item 40), much less than the .20 average correlation from previous studies. Expected grade correlated highest with ratings of Assignments, Exams and Grading (.17), Course Outcomes (.16), and Course Difficulty/Workload (.17). Course Difficulty/Workload, the other primary independent variable (and for this analysis a mean of the three items in Table 1), correlated −.53 with Student Effort and Involvement, indicating that students put more effort into courses they rated as more difficult and as having a heavier workload. Course Difficulty/Workload correlated .30 with Assignments, Exams, and Grading, meaning courses seen as less difficult and with lighter workloads were rated as more effective in such areas as the clarity and appropriateness of exams, grades, and assignments. However, Difficulty/Workload correlated only .06 with Overall Evaluation and only a little higher with the other instructional scales. Other high correlations in Table 3 confirmed expectations: upper-level courses were electives or in a major/minor (.43), and upper-level courses were prevalent at 4-year rather than 2-year colleges (.41).

Table 4 provides the correlations of key variables within each of the eight subject areas. Examining these results indicates that there were sizeable differences among the subject areas on many of the variables. Expected grades, for example, correlated with Overall Evaluation .15 in Natural Sciences and .14 in Business, compared with .05 and .06 in Education and Fine Arts. With Course Outcomes, expected grade correlated from .20 in Natural Sciences to .03 in Health. Difficulty/Workload correlated with Overall Evaluation 0.13 (Fine Arts), −.06 (Education), and −.01 (Humanities), compared to .14 (Natural Sciences) and .15 (Health). With Course Outcomes, Difficulty/Workload correlated −.22 (Fine Arts), −.20 (Health), −.19 (Education), and −.19 (Engineering and Technology), compared with positive values of .09 in Business, .06 in Natural Science, and .05 in Social Science. The eight subject areas also varied in their means and standard deviations, as shown in Table 5. For the 14 variables, 5 had at least a standard deviation difference in their means, and another 6 varied by about a half standard deviation. Of the two primary independent variables, Difficulty/Workload varied from a mean of 2.90 in Education to 2.51 in Health, about one standard deviation difference; expected grade varied from 4.87 in

TABLE 3. Correlation Matrix (N = 46,687–55,155)

	A	B	C	D	F	G	H	40	41	Cl Siz	In Typ	Cl Lev	Ex Gr	Lec
Scales[a]														
A														
B	92													
C	80	83												
D	85	84	82											
F	76	78	70	78										
G	44	43	32	52	64									
H	07	10	19	30	02	–53								
Ov. Eval. 40	89	88	80	83	82	47	06							
Crs. Req. 41	06	10	07	03	21	17	–14	11						
Cl. Size	07	09	11	10	17	16	–05	09	21					
In. Type	–10	–09	–09	–19	–16	–21	00	–08	13	–08				
Cl. Level	–02	03	01	07	05	00	–05	02	43	13	41			
Exp. Grade	10	12	15	17	16	02	17	11	04	05	–03	05		
Lecture/Disc.	01	02	02	02	–06	–08	05	00	–09	–10	05	07	00	
Disc./Lab.	–01	01	02	01	13	11	–02	02	09	16	–09	–05	04	–82

[a]A = Course Organization and Planning. B = Communication. C = Faculty/Student Interaction. D = Assignments, Exams and Grading. F = Course Outcomes. G = Student Effort and Involvement. H = Course Difficulty/Workload.

TABLE 4. Correlations of Expected Grade and Difficulty/Workload (D/W) with Key Variables, by Eight Subject Areas

	Health $N = 2,465$		Business $N = 5,446$		Education $N = 3,693$		Social Science $N = 9,787$		Fine Arts $N = 3,171$		Natural Science $N = 10,590$		Eng. and Tech. $N = 6,397$		Humanities $N = 12,943$	
Scales[a]	Ex. Gr.	D/W	Ex. Gr.	D/W	Ex. Gr.	D/W	Ex. Gr.	D/W	Ex. Gr.	D/W	Ex. Gr.	D/W	Ex. Gr.	D/W	Ex. Gr.	D/W
A	11	07	.13	11	06	02	09	06	06	–05	13	12	12	08	10	00
B	13	09	15	14	07	00	11	09	07	–05	15	15	13	09	12	03
C	16	19	18	22	08	09	17	22	07	07	17	23	16	22	14	13
D	15	16	20	21	08	09	19	24	09	02	21	28	18	21	15	14
F	03	–20	19	09	08	–19	16	05	07	–22	20	06	14	– 19	15	00
G	–11	–65	01	–51	07	–57	03	–49	05	–54	–04	–60	05	–50	05	–43
Overall Eval.	08	15	14	12	05	–06	11	07	06	–13	15	14	11	06	11	–01
Exp. Grade	1.00	27	1.00	23	1.00	01	1.00	18	1.00	05	1.00	23	1.00	17	1.00	14
Diff/Work H	27	1.00	23	1.00	01	1.00	18	1.00	05	1.00	23	1.00	17	1.00	14	1.00

[a]A = Course Organization and Planning. B = Communication. C = Faculty/Student Interaction. D = Assignments, Exams and Grading. F = Course Outcomes. G = Student Effort and Involvement. H = Course Difficulty/Workload.

TABLE 5. Means and Standard Deviations of Key Variables by Eight Subject Areas[a]

	Health *N* = 2,465		Business *N* = 5,646		Education *N* = 3,693		Social Science *N* = 9,787		Fine Arts *N* = 3,171		Natural Science *N* = 10,590		Eng. and Tech N = 6397		Humanities N = 12,943	
	X	S.D.	X	S.D.	X	S.D.	X	S.D.	X	S.D.	X	S.D.	X	S.D.	X	S.D.
Scale G: Student Effort Inv.	3.93	.47	3.68	.39	3.67	.48	3.63	.38	3.70	.48	3.73	.39	3.69	.42	3.68	.39
Coll. Req. vs. Choice: Item 41	.84	.19	.75	.25	.76	.28	.65	.30	.73	.29	.59	.32	.73	.25	.44	.33
Class Size/Small	.42	.49	.33	.47	.31	.46	.24	.43	.49	.50	.29	.46	.43	.50	.29	.45
Institutional Type	.35	.48	.70	.46	.81	.39	.73	.45	.72	.45	.68	.47	.53	.50	.73	.44
Teacher: Lecture/Disc.	.53	.50	.73	.44	.59	.49	.78	.42	.42	.49	.44	.50	.28	.45	.73	.45
Teaching: Lab./Disc.	.41	.49	.18	.39	.38	.49	.10	.30	.50	.50	.38	.49	.67	.47	.24	.43
Scale H: Crse. Diff., Work, Pace	2.51	.43	2.67	.34	2.90	.36	2.75	.30	2.81	.33	2.59	.36	2.72	.36	2.73	.30
Expected Grade	4.62	1.28	4.62	1.17	4.87	1.81	4.50	1.13	4.59	1.53	4.33	1.00	4.60	1.35	4.51	1.16
Overall Eval.: Item 40	4.12	.50	3.96	.54	4.11	.49	4.06	.48	4.09	.49	3.95	.53	3.93	.53	4.05	.49
Scale F: Course Outcomes	3.65	.50	3.65	.50	3.88	.48	3.71	.47	3.88	.49	3.55	.50	3.76	.50	3.71	.48
Scale A: Org. and Plan.	4.34	.48	4.23	.50	4.37	.44	4.32	.42	4.31	.44	4.24	.46	4.16	.49	4.30	.43
Scale B: Commun.	4.40	.43	4.25	.48	4.44	.39	4.35	.41	4.39	.39	4.24	.46	4.19	.47	4.35	.41
Scale C: F/S Int.	4.37	.51	4.27	.51	4.44	.43	4.32	.45	4.35	.47	4.27	.49	4.23	.51	4.34	.47
Scale D: As.,Ex.,Gr.	4.17	.47	4.06	.47	4.22	.44	4.08	.43	4.15	.43	4.04	.44	4.02	.47	4.14	.43

[a]See Table 2 for explanation of responses.

Education to 4.33 in Natural Science, slightly less than half a deviation difference. The differences in the correlations and means for the eight subject areas support the inclusion of subject area as a variable in the analyses.

Multiple Regression Results

Table 6 lists the standardized β weights for each variable. Because of the large sample sizes, many of the β weights are significant at the .01 level in spite of being small and of little practical value. Among the controlled variables, Course Outcomes has the largest β weights, ranging from .79 to .96 for Scales A through D and Overall Evaluation. The β weights for Student Effort and Involvement and for teaching by discussion or in labs were next highest in size but considerably lower than Course Outcomes. Their negative values may be due to interaction with other controlled variables such as class size.

The β weights for Difficulty/Workload and expected grade at the bottom of Table 6 are noteworthy. They apply to all eight of the subject areas. In general, they indicate that the level of difficulty, workload, and pace in a course has a greater influence on the dependent variables than do expected grades. The linear values are positive (.66–.37) and the quadratic values are negative (–.30––.56), suggesting that courses get higher ratings as they go from being too difficult to about right, but that when they are rated somewhat elementary or having a lighter workload and pace, they are rated slightly lower.

There are, however, differences among the eight subject areas. The best way to interpret subject area results is by inspecting their β weights in Table 6 along with the predicted values for each of the dependent variables depicted as bar graph figures. Because there were 40 of these figures (five dependent variables for each of the eight subject areas), they are included only in Centra (2002). As examples, Figs. 1, 2, and 3 are included here; they represent the prediction of the Overall Evaluation item for Business, Social Science, and Natural Science. The effects of the eight control variables (Table 6) have been accounted for in the expected grade and Difficulty/Workload predictors in the figures. While the scale for grades covers the full 7-point range of A to below C, the scale for Difficulty/Workload only runs from 1.50 to 3.50, that is, from between very difficult and somewhat difficult, to between about right and somewhat elementary. This abbreviated range was necessary because, as Table 1 shows, there were few responses at the extreme elementary (lighter, slow) end of the scale, and this was especially true for the small subject areas. Results for each subject area follow.

Business

For Overall Evaluation and each of the scales, the lowest evaluations were given by students who rated courses as most difficult, a result for the other

TABLE 6. Stepwise Multiple Regression of Dependent Variables (Four Scales and Overall Evaluation) on Subject Area, Course Difficulty/Workload, and Expected Grade, Controlling for Course Outcomes and Other Selected Variables (N = 46,687–55,155)[a]

	Overall Eval. N = 53,549	A Crse. Org. and Planning N = 53,515	B Communi-cation N = 53,388	C Fac/St Interaction N = 53,297	D Assign, Ex., Grad'g N = 53,297
Controlled Variables					
Student Effort/ Involvement (Scale G)	−.15	−.09	−.13	−.15	−.01
Coll. Req. Crse./vs. Choice	−.03	−.07	−.03	−.04	−.06
Class Size	−.01	−.01	−.01	.03	ns
Institutional Type	.02	.01	.01	ns	−.05
Class Level	−.02	−.03	.01	−.01	−.04
Teaching: Lecture/ Disc.	−.06	−.09	−.04	−.01	−.05
Teaching: Disc./Lab.	−.11	−.16	−.10	−.07	−.11
Course Outcomes (Scale F)	.96	.89	.91	.82	.79
Other Variables					
Business	.08	ns	ns	.12	.11
Business × Diff./Work (linear)	ns	.71	.26	ns	ns
Business × Diff./Work (Quad)	ns	−.65	−.25	−.09	−.07
Business × Grade	ns	−.28	ns	ns	ns
Business × Diff./Work (linear) × Grade	ns	.24	ns	ns	ns
Business × Diff./Work (Quad) × Grade	−.04	ns	ns	ns	ns
Education	−.03	ns	ns	ns	ns
Education × Diff./ Work (Quad)	.03	ns	ns	ns	ns
Social Sciences	.16	.39	.29	.16	ns
Soc. Sci. × Diff./Work (linear)	ns	ns	ns	−.14	ns
Soc. Sci. × Diff./Work (Quad)	−.06	−.37	−.29	ns	ns
Soc. Sci. × Diff./Work (Quad) × Grade	−.03	.17	.13	ns	ns
Soc. Sci. × Grade (Quad)	ns	−.14	−.10	ns	.01
Fine Arts	ns	ns	ns	−.03	ns

TABLE 6. (*continued*)

	Overall Eval. $N = 53,549$	A Crse. Org. and Planning $N = 53,515$	B Communi-cation N = 53,388	C Fac/St Interaction $N = 53,297$	D Assign, Ex., Grad'g $N = 53,297$
Fine Arts × Diff./Work (Quad)	ns	–.02	ns	ns	ns
Natural Sciences	.36	.66	.48	.19	–.12
Nat. Sci. × Diff./Work (linear)	–.29	–.60	–.44	–.22	ns
Nat. Sci. × Diff./Work (linear) × Grade	.20	.28	.20	ns	ns
Nat. Sci. × Grade (linear)	ns	ns	ns	.43	.61
Nat. Sci. × Grade (Quad)	–.15	–.21	–.14	–.28	–.37
Tech.	–.02	ns	ns	–.04	–.04
Tech. × Diff/Work (linear)	ns	.08	ns	ns	ns
Tech. × Diff/Work (Quad)	ns	–.13	–.10	ns	ns
Tech. × Diff/Work (Quad) × Grade	ns	ns	.03	ns	ns
Humanities	.15	.33	.26	.32	.29
Humanities × Diff./ Work (linear)	ns	ns	ns	–.27	–.24
Humanities × Diff./ Work (Quad)	–.14	–.36	–.21	ns	ns
Humanities × Diff/ Work × Grade	ns	.22	ns	ns	ns
Humanities × Grade (linear)	.16	ns	ns	ns	ns
Humanities × Grade (Quad)	–.12	–.16	ns	ns	ns
Diff./Work (linear)	.57	.49	.37	.66	.65
Diff./Work (Quad)	–.56	–.39	–.30	–.52	–.49
Exp. Grade (linear)	ns	ns	ns	–.22	–.37
Exp. Grade (Quad)	ns	ns	ns	.25	.41
R^2	.72	.64	.67	.56	.68

[a]Standardized Beta Weights: All t values were significant at .01 level unless indicated by ns or excluded.

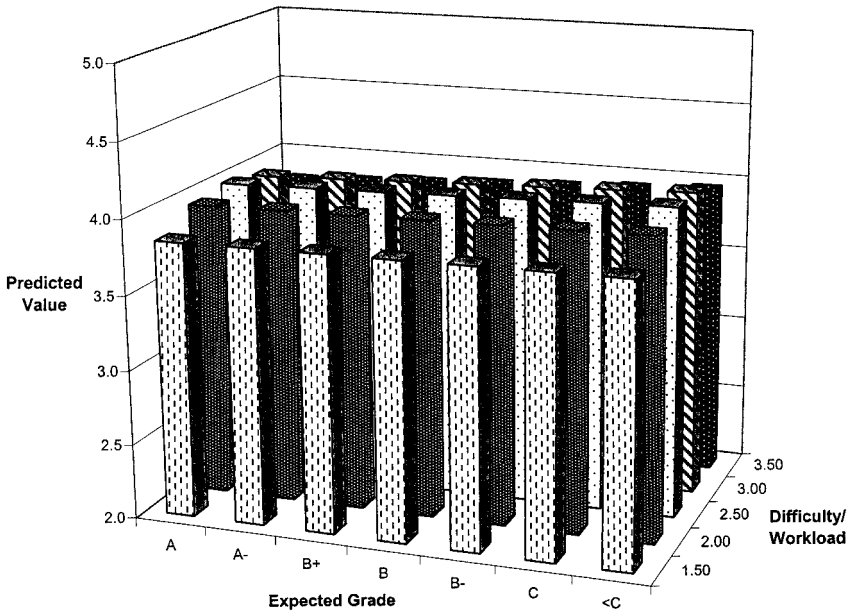

FIG. 1. Prediction of overall evaluation for business.

subject areas as well. A steady linear increase to 3.00, about right, is evident on all five dependent variables, and with the exception of Scales C and D, a drop in ratings at 3.50. Scale A, as the β weights suggest, had the most dramatic drop at 3.50, and also an interaction in that courses with A grades and higher difficulty had the lowest evaluations. Grades were generally unrelated to ratings except for Scale D, in which A courses received higher ratings.

Social Sciences

For Difficulty/Workload, evaluations for Overall Evaluation and Scales A and B increased to 3.00 and then decreased at 3.50. For Scales C and D there was no decrease at 3.50, and for these two scales courses with higher average grades received higher ratings. Otherwise, grades showed no relationship to evaluations (Overall Evaluation), or higher grades resulted in slightly lower course evaluations for courses averaging an A (Scales A and B).

Natural Science

The results for Difficulty/Workload differed from other fields in that the relationships between Difficulty/Workload and evaluations were much smaller for

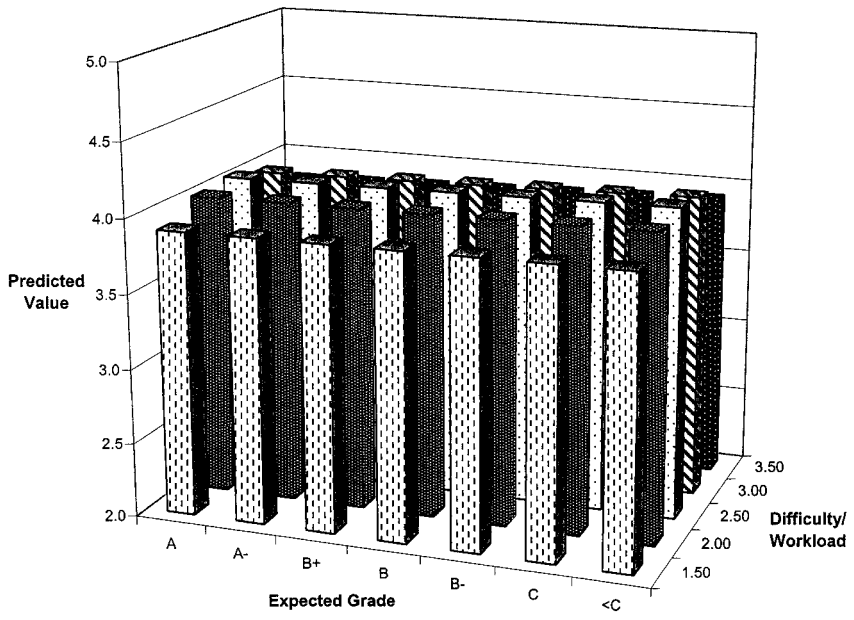

FIG. 2. Prediction of overall evaluation for social science.

Overall Evaluation and Scales A and B (β weights cancelled those at the bottom of Table 6). For Scales C and D, a stronger relationship was found, but unlike most other subject areas, ratings did not decrease at 3.50. It may be that few courses existed at that level because few were rated easy. Grades showed no relationship to evaluations except for a small decrease in Overall Evaluations and Scale A for A graded courses rated as difficult.

Humanities

On the Difficulty/Workload dimension, evaluations for Overall Evaluation and Scale A increased to 3.00 and then decreased considerably at 3.50; Scales B and C did not decrease at 3.50. Grades were either flat or, for Scales C and D, curvilinear in that courses with expected grades of A or C gave slightly higher evaluations than those with B grades.

Engineering and Technology, Education, Fine Arts, and Health

For all four of these subject areas, Difficulty/Workload was linearly related with evaluations of courses on all scales, with lowest evaluations for those rated most difficult/heaviest workload. Only for Overall Evaluation was there a slight

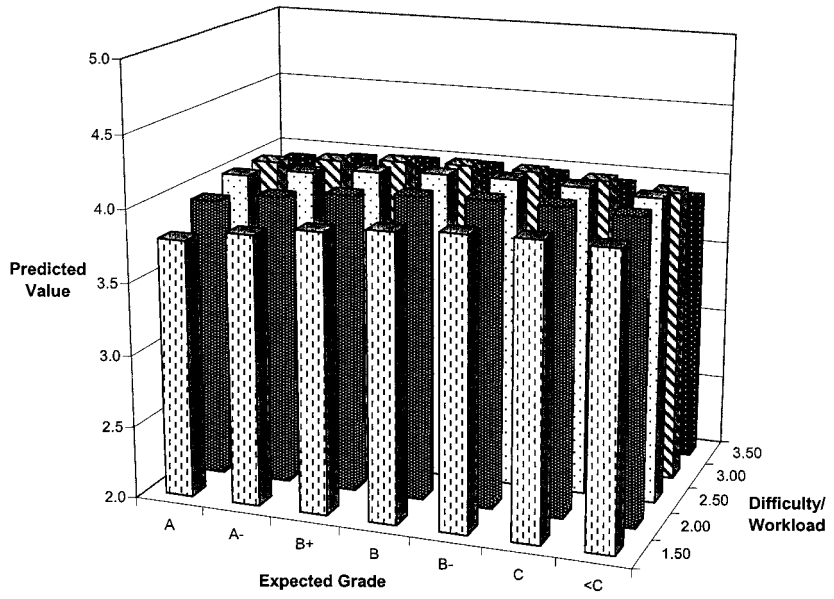

FIG. 3. Prediction of overall evaluation for natural science.

decrease in evaluations when the course was seen as slightly elementary or lighter. For expected grade, there was no relationship for the first three dependent variables and a small curvilinear relationship for Scales C and D (that is, courses with average grades of A or C were rated higher than those with B grades).

DISCUSSION

The average expected grade in courses correlated only .11 with the overall evaluation question used in this study, compared with a .20 correlation average from other studies. This lower correlation may be due to the particular wording in the SIR II questionnaire. While many other forms use a global question that asks students to rate the teacher or the course, generally on an excellent to poor scale, the SIR II global question asks students to rate the quality of instruction as it contributed to their learning in the course (very effective to ineffective). Most of the other statements relating to specific instructional practices are given the same emphasis—how each contributed to learning. This emphasis probably takes the focus away from students' general liking of the teacher or the course and places it on their perceptions of what they have learned.

Correlations of expected grades and course evaluations were a little higher

for two of the scales (.17, Scale D and .15, Scale C) and were slightly higher for some subject areas (Natural Science and Business). Results of the regression analyses, however, demonstrated the minimal effect of expected grades on course evaluations. The regression analyses, with the resulting β weights and predicted values, controlled for student self-reports of learning through the Course Outcomes Scale, as well as for other variables that might affect student evaluations of courses. As the predicted evaluation values illustrate in Figs. 1 through 3, for the eight subject areas, there was clearly no relationship in 27 of 40 cases; the predicted values as reflected by the height of the bars were the same for A, B, and C graded courses, and this was true for the various levels of difficulty of courses as well. For the smaller subject areas, Health, Education, and Fine Arts, and in Engineering/Technology, Social Science, Business, and Humanities, A and C graded courses were rated higher than B courses on Scales C and D, a curvilinear relationship. In no instances, and this includes the eight subject areas and five course evaluations, did courses with A level expected grades receive higher evaluations. In fact, in Natural Science courses with A level expected grades were rated a little lower on three evaluation measures, especially if the courses were rated as difficult.

While the average expected grade instructors had given in their courses had little affect on the student evaluations of those courses, the findings for Difficulty/Workload were more complex. Students rated most courses "about right" on the Difficulty/Workload scale. Moreover, courses were about four times more likely to be rated at the difficult/heavy/fast end of the scale than the elementary/lighter/slow end (Table 1). The question raised in this study is whether those courses rated as easier, even though they are in the minority, also received higher student evaluations due to a student bias for such courses. The correlations in Tables 3 and 4 indicate a modest relationship between Difficulty/Workload and evaluations of courses, but these are linear correlations. A plot of the values supported what Table 1 suggests: the relationship was curvilinear, thus calling for a quadratic as well as linear examination of the Difficulty/Workload measure. The question of bias, however, can only be investigated by controlling for other variables that may affect course evaluations, such as the teaching method, class size and, especially, student learning (the Course Outcomes scale). A review of the β weights and the predicted evaluation values from the regression analyses indicate that courses seen as difficult were always rated lowest. In all of the eight subject areas, evaluations on all measures increased to the 3.0 midpoint ("about right") or 3.5 level of the Difficulty/Workload measure. Due to the few courses at the easy end of the scale, the figures included predicted values to only the 3.5 point of the 5-point scale, which is just beyond the "about right" midpoint. Slightly more than half of the 40 predicted evaluations in the figures peaked at the 3.5 point; the remaining evaluations peaked at the 2.5 or 3.0 level and then dropped. In these latter instances, courses were thus rated

lower when they were seen as somewhat elementary or slow, which is contrary to what some faculty members believe. What these findings indicate is that teachers will receive better evaluations when their courses are manageable for students. In other words, students will view instruction as most effective when it is at their level of preparation and ability rather than too difficult, when the course workload is close to what other courses demand rather than much heavier, and when the pace at which material is covered is about right for them rather than too fast. All of this makes sense for good instructional design; teachers should be aware of their students' ability level and preparation when presenting material and giving assignments. A similar conclusion was reached by Marsh and Roche (2000), even though their definition of workload differed from this study's definition of Difficulty/Workload.

Subject area differences are noteworthy. Natural Science courses tended to be rated among the most difficult while giving the lowest average grades (Table 5). Also in Natural Science, on the Overall Evaluation question and on Scales A and B the evaluations students gave in difficult courses were not much lower than evaluations in less difficult courses. As noted earlier, the lowest evaluations were given in A graded courses seen as difficult. Courses with average expected grades of C and B rated the course similarly regardless of its difficulty/workload level. While this pattern did not repeat itself for Scales C and D, it did occur on Scale A in Business, Social Science, and Humanities, indicating that high achieving students can be especially critical of courses they see as having a high level of difficulty/workload. Scale A, Course Organization and Planning, contains items on the instructor's preparation for class and command of subject matter, areas in which high achieving students may have higher expectations of teachers.

A comparison among the four evaluation scales clearly indicates that Scales C and D are most influenced by the level of course difficulty. In all eight-subject areas, courses rated the most difficult were given evaluations on Scales C and D well below those that were less difficult, with a sharp incline in those evaluations to the 3.50 level. The reasons Faculty/Student Interaction (Scale C) and Assignments, Exams, and Grading (Scale D) would be especially responsive to course difficulty are not entirely evident. Faculty/Student Interaction includes ratings of an instructor's helpfulness and concern for students' learning. The Assignments, Exams, and Grading category includes measures of exam fairness and the helpfulness of assignments. These are areas that are important to students for their learning and for which courses seem to vary greatly in difficulty.

By statistically controlling for student self-reported learning (Course Outcomes, Scale F), this study was better able to investigate bias in student evaluations due to grading leniency or course workload. According to the definition of bias used in this study, correlations of expected grades with course evaluations are due in part to validity. Students who learn more in a course expect to get higher grades and also believe instruction has been more effective. The high-

standardized β weights for self-reported learning did, in fact, attest to its importance in determining student evaluations. All other variables controlled were relatively minor in influence, with the exception of student Effort and Involvement (Scale G). Bias due to grades or workload was generally nonexistent, a finding that coincided with Marsh and Roche's (2000) path-analytic study based on 12 years of data at one institution. The study reported here found little evidence of bias in eight different subject areas as well. In spite of lower grades and lower student evaluations in Natural Science courses, no evidence of a grading leniency or workload bias existed even in those courses. In fact, students with higher expected grades gave somewhat lower evaluations, just the opposite of a grading leniency expectation.

To summarize, teachers will not likely improve their evaluations from students by giving higher grades and less course work. They will, however, improve their evaluations and probably their instruction if they respond to consistent student feedback about instructional practices (Centra, 1993).

ACKNOWLEDGMENT

Tim Wasserman provided help with the analyses, I thank him.

REFERENCES

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

Baird, L. L. (1976). *Using Self-Reports to Predict Student Performance*, The College Board, New York.

Cashin, W. E. (1988). Student ratings of teaching: A summary of research. IDEA Report No. 20, Kansas State University, Division of Continuing Education.

Cashin, W. E. (1990). Students do rate different academic fields differently. In: Theall, M., and Franklin, J. (eds.), *Student Ratings of Instruction: Issues for Improving Instruction*, New Directions for Teaching and Learning, No. 43, Jossey-Bass, San Francisco.

Centra, J. A. (1972). *The Student Instructional Report: Its Development and Uses*, Educational Testing Service, Princeton, NJ.

Centra, J. A. (1998). *Development of the Student Instructional Report II,* Educational Testing Service, Princeton, NJ.

Centra, J. A. (1993). *Reflective Faculty Evaluation*, Jossey-Bass, San Francisco.

Centra, J. A. (2002). Will teachers receive higher student evaluations by giving higher grades and less coursework. Research Report No. 10, The Student Instructional Report II, Educational Testing Service, Princeton, NJ.

Centra, J. A., and Creech, F. R. (1976). The relationship between student, teachers, and course characteristics and student ratings of teacher effectiveness. Project Report 76-1, Educational Testing Service, Princeton, NJ.

Centra, J. A., and Gaubatz, N. B. (2000a). Is there gender bias in student evaluations of teaching? *J. Higher Educ.* **71:** 17–33.
Centra, J. A., and Gaubatz, N. B. (2000b). Student perceptions of learning and instructional effectiveness in college courses. Research Report No. 9, The Student Instructional Report II, The Educational Testing Service, Princeton, NJ.
d'Apollonia, S., and Abrami, P. C. (1997). Navigating student ratings of instruction. *Am. Psychol.* **52:** 1198–1208.
Dowell, D. A., and Neal, J. A. (1982). A selective view of the validity of student ratings of teaching. *J. Higher Educ.* **53:** 51–62.
Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers and courses: What we know and what we don't. *Res. Higher Educ.* **9:** 199–242.
Feldman, K. A. (1989). Association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Res. Higher Educ.* **30:** 583–645.
Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In: Perry, R. P., and Smart, J. C. (eds.), *Effective Teaching in Higher Education: Research and Practice*, Agathon, New York, pp. 368–395.
Franklin, J., and Theall, M. (1996). Disciplinary differences in sources of systematic variation in student ratings of instructor effectiveness and students' perceptions of the value of class preparation time: A comparison of two universities' ratings data. Paper presented at the annual meeting of the American Educational Research Association, New York.
Franklin, J., Theall, M., and Ludlow, L. (1991). Grade inflation and student ratings: A closer look. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
Gillmore, G. M., and Greenwald, A. G. (1994). *The Effects of Course Demands and Grading Leniency on Student Ratings of Instruction*, Office of Educational Assessment (94-4), University of Washington, Seattle.
Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *Am. Psychol.* **35:** 603–618.
Greenwald, A. G., and Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *J. Educ. Psychol.* **89:** 743–751.
Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *J. Educ. Psychol.* **63:** 130–133.
Howard, G. S., and Maxwell, S. E. (1980). Correlation between student satisfaction and grades: A case of mistaken causation? *J. Educ. Psychol.* **72:** 810–820.
Koon, J., and Murray, H. S. (1996). Using multiple outcomes to validate student ratings of overall teacher effectiveness. *J. Higher Educ.* **66:** 61–81.
Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *Int. J. Educ. Res.* **11:** 253–288.
Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on student evaluations of teaching. *Am. Educ. Res. J.* **38:** 183–212.
Marsh, H. W., and Roche, L. A. (1997). Making students evaluations of teaching effectiveness effective. *Am. Psychol.* **52:** 1187–1197.
Marsh, H. W., and Roche, L. A. (2000). Effects of grading leniency and low workloads on students' evaluations of teaching: Popular myth, bias, validity or innocent bystanders? *J. Educ. Psychol.* **92:** 202–208.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe* **65:** 384–397.
Neter, J., Kutner, N. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Regression Models* (3rd Ed.), Irwin, Chicago.
Pike, G. R. (1995). The relationship between self-reports of college experiences and test scores. *Res. Higher Educ.* **36:** 1–22.
Powell, R. W. (1977). Grades, learning, and student evaluation of instruction. *Res. Higher Educ.* **7:** 193–205.
Vasta, R., and Sarmiento, R. F. (1979). Liberal grading improves evaluations but not performance. *J. Educ. Psychol.* **71:** 207–211.
Worthington, A. G., and Wong, P. T. P. (1979). Effects of earned and assigned grades of student evaluations of an instructor. *J. Educ. Psychol.* **71:** 764–775.

Received June 7, 2002.