ETS SIR II™

Listening. Learning. Leading.®

Will Teachers Receive Higher Student Evaluations By Giving Higher Grades and Less Coursework?

John A. Centra

July 9, 2005

Will Teachers Receive Higher Student Evaluations By Giving Higher Grades and Less Coursework?

John A. Centra

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 student evaluations of teaching in tenure and promotion decisions at many colleges, a teacher's temptation to manipulate grades or course workload would be understandable.

Certainly no method of evaluating college teaching has been researched more than student evaluations, with well over 2,000 studies referenced in the ERIC system. The majority of these study results have been positive, concluding that the evaluations are (1) reliable and stable; (2) valid when compared to student learning and other indicators of effective teaching; (3) multidimensional, in terms of what they assess; (4) useful in improving teaching; and (5) only minimally affected by various course, teacher, or student characteristics that could bias results (Cashin, 1988;

Centra, 1993; d'Appolonia & Abrami, 1997; Marsh, 1987; McKeachie, 1979). Recently, several studies have addressed the effects of grading leniency and course workload on student evaluations, spurred by the strong opinions held by many faculty members. 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. This study will use 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 only of the grades 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 actual final grades (Franklin, Theall, & Ludlow, 1991).

Expected grades are correlated moderately with student evaluations: In a study of 9,194 class-average Student Instructional Report responses from a variety of colleges and courses, expected grades and global ratings of teacher effectiveness correlated .20 (Centra & 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 the instructor or the course is rated, 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 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 will be 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 courses and the instruction they have received (Howard & 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, but low grades (and the ratings students give) would be attributed to poor instruction.

Experimental field studies that investigated a grading-leniency bias on ratings used a design in which students were given false course grades (Holmes, 1972; Powell, 1977; Vasta & Sarmiento, 1979; Worthington Wong, 1979. Although 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 it 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 those of students in the class performing at the same level, was both unreasonable and offensive to students. Moreover, 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 will investigate 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 were also included in a "workload" factor. Gillmore and Greenwald (1994) included hours per week along with three other items to infer workload: the students' reported challenge of the course, their effort, and their involvement. This definition of workload would seem 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 data base demonstrated clearly that a student effort/involvement/challenge factor was highly correlated with evaluations of instruction (Centra & 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, that a lower workload was related to higher student evaluations (and that most teachers believe to be true), was reported by Franklin, Theall, and Ludlow (1991), although the effect size was small. The difference among these findings is likely due to the different definitions of workload: Franklin, Theall, and Ludlow did not include hours students reported studying outside of class, whereas the previous two studies did. Hours spent outside of class on coursework can be further refined by dividing then 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 and students' perceptions of their learning, whereas 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 (e.g., Centra & 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 correlate with student evaluations but are not necessarily biasing effects (Centra & 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 do more experienced teachers, but because students often 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 Gaubatz, 2000a).

Applying this definition of bias to this study, it is important to investigate the relationships of expected grades and difficulty/workload not only to course evaluations but also to a measure of student learning in the course. Although this study will not have available an objective measure of learning, such as a 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 & Neal, 1982; Feldman, 1989; Koon & 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 will control for student self-reported learning outcomes. In addition, unlike in previous studies, the analyses will 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 Educational Testing Service in the early 1970s (Centra, 1972). The SIR II has some of the same instructional evaluation scales as the earlier SIR 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 (1988).

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 other courses" for course workload. Also, unlike in most other rating instruments, each of the five points on the scale is described, the mid/point of 3 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 mid/point ("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 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 alpha and test-retest reliabilities (Centra, 1998). The Course Organization and Planning scale (Scale A, five items) included the instructor's explanation of course requirements, use of class time, and emphasis of important points. The Communication scale (Scale B, five items) included the instructor's ability to make clear presentations, use challenging questions or problems, and be enthusiastic about the course material. Within the Faculty/Student Interaction scale (Scale C, five 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, six 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 contributed to their learning. In short, the emphasis of the form is in tying practices to learning, and in making students aware of and responsive to that connection. And whereas 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 students in what they learned.

A number of variables in addition to the primary ones were also used as independent or control variables (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 & 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 years, 1=4 year or more

Class Size
0=16 or larger, 1=6 to 15

Class Level
0=freshman/sophomore
1=junior/senior

College-Required Course vs Choice
0=college required
1=major/minor requirement 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

Course Outcomes Scale
5=high
1=low.

Sample

The sample for this study included a 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 and 55,549 classes. As Table 2 indicates, about 32 % of these classes were in two-year colleges and 68 % were in four-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 a B+; the lecture teaching method was dominant, and the overall evaluation (item 40) as well as Scales A-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 rather than the individual student class was the unit of analysis. 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 the evidence of some curvilinearity in the responses. Expected grade was also entered as both a linear and a quadratic function because of 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. Because of the large N in this study multicollinearity was not a problem, although not surprisingly some variables were highly correlated (Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. 1996).

Results

As the correlation matrix in Table 3 indicates several of the SIR II scales, together with the overall evaluation item (#40), are highly intercorrelated. Because these are correlations of mean values, they are much higher than individual scores 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, 1998). 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 2), 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 2 confirmed expectations: upper-level courses were electives or in a major/minor (.43), and upper-level courses were prevalent at four-year rather than two-year colleges (.41).

Table 4 provides the correlations of key variables within each of the eight subject areas. Examining these results indicate that there were sizable 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 to .05 and .06 in education and fine arts. With course outcomes, expected grades 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), com-pared 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 to 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 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 beta weights for each variable. Because of the large sample sizes, many of the beta weights are significant at the .01 level in spite of being small and of little practical value. Among the controlled variables, course outcomes had the largest beta weights, ranging from .79 to .96 for Scales A-D and overall evaluation. The beta weights for student effort and involvement and for teaching by discussion or in labs were next highest in size but considerably lower than those for course outcomes. Their negative values may be because of interaction with other controlled variables such as class size.

The beta 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 to .37) and the quadratic values are negative (-.30 to -.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 as having a lighter workload and pace, they are rated slightly lower.

There are, however, differences among the eight subject areas. Inspecting their beta weights in Table 6 along with the predicted values for each of the dependent variables in Figures 1, a-d through Figures 8 a-d, is the best way to interpret the subject area results. Because there are 40 of these figures (five dependent variables for each of the eight subject areas), they are included in the Appendix. The effects of the eight control variables (Table 6) have been accounted for in the expected grade and difficulty/workload predictors in the figures. Whereas the scale for grades covers the full seven point range of A to below C, the scale for difficulty/workload runs only 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 (Figure 1 a-d). 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 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 beta 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 (Figure 2 a-d). 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 (Figure 3 a-d). The results for difficulty/workload differed from those for other fields in that the relationships between difficulty/workload and evaluations were much smaller for overall evaluation and Scales A and B (beta weights canceled those at the bottom of Table 6). For Scales C and D a stronger relationship was found but unlike in 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 (Figure 4 a-d). 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 (Figure 5 a-d), Education (Figure 6 a-d), Fine Arts (Figure 7 a-d), Health (Figure 8 a-d). 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 decrease in evaluations when the course was seen as slightly elementary or lighter. For expected grade, there was no relationship for the first three of the 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 to 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 to 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 beta 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 Figures 1-8, for the eight subject areas there was clearly no relationship in 27 of 40 cases; that is, 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.

Although the average expected grade instructors had given in their courses had little effect 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 at 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 because of 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, that the relationship was curvilinear, thus calling for a quadratic as well as a linear examination of the difficulty/workload measure. The question of bias, however, can only be investigated by controlling for other variables that may effect course evaluations, such as the teaching method, class size, and, especially, student learning (the Course Outcomes scale). A review of the beta weights and the predicted evaluation values from the regression analyses indicates that courses seen as difficult were always rated lowest. In all eight subject areas, evaluations on all measures increased to the 3.0 midpoint (about right) or 3.5 level of the difficulty/workload measure (Figures 1-8). Because of the few courses at the easy end of the scale, Figures 1-8 includes 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 Figures 1-8 peaked at the 3.5 point, and 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 many 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 the students 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. Although 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 indicates clearly 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, and assignments, exams, and grading 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 because of 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; i.e., students who learn more in a course expect to get higher grades and also believe instruction has been more effective. The high standardized beta 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 non existent, 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).

References

Baird, L. L. (1976). Using self-reports to predict student performance. New York: The College Board.

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

Cashin, W. E. (1990). "Students do rate different academic fields differently." In M. Theall and J. Franklin (Eds.) Student ratings of instruction: issues for improving practice. New Directions for Teaching and Learning, No. 43, San Francisco, Jossey-Bass.

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

Centra, J. A. (1988). Development of the Student Instructional Report II. Princeton, NJ: Educational Testing Service.

Centra, J. A. (1993). Reflective faculty evaluation. San Francisco: Jossey-Bass.

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

Centra, J. A. & Gaubatz, N. B. (2000a). Is there gender bias in student evaluations of teaching? Journal of Higher Education, 71, 17-33.

Centra, J. A. & Gaubatz, N. B. (2000b). Student perceptions of learning and instructional effectiveness in college courses (Research Rep., No. 9, The Sudent Instructional Report II) Princeton, NJ Educational Testing Service.

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

Dowell, D. A. & Neal J. A. (1982). A selective view of the validity of student ratings of teaching. Journal of Higher Education, 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. Research in Higher Education, 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. Research in Higher Education, 30, 583-645.

Feldman, K. A. (1997). Identifying exemplary tachers and taching: Edence from student ratings. In R. P. Perry and J. C. Smart (Eds.), Effective Teaching in higher education: Research and practice (pp. 368-395). New York: Agathon.

Franklin, J., & 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., & 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., & 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. American Psychologist, 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. Journal of Educational Psychology, 89, 743-741.

Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. Journal of Educational Psychology, 63, 130-133.

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.

Koon, J., & Murray, H. S. (1996). Using multiple outcomes to validate student ratings of overall teacher effectiveness. Journal of Higher Education, 66, 61-81.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research, 11, 253-288.

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on student evaluations of teaching. American Educational Research Journal, 38, 183-212.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. American Psychologist, 52, 1187-1197.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workloads on students' evaluations of teaching: Popular myth, bias, validity or innocent bystanders? Journal of Educational Psychology, 92, 202-208.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. Academe, 65, 384-397.

Neter, J., Kutner, M. H., Nachtsheim, C. J. Wasserman, W. (1996) Applied linear regression models (3rd ed). Chicago: Irwin.

Pike, G. R. (1995). The relationship between self-reports of college experiences and test scores. Journal of Research in Higher Education, 36, 1-21.

Powell, R. W. (1977). Grades, learning, and student evaluation of instruction. Research in Higher Education, 7, 193-205.

Vasta, R., & Sarmiento, R. F. (1979). Liberal grading improves evaluations but not performance. Journal of Educational Psychology, 71, 207-211.

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

TABLE 1

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

For my preparation and ability, the 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

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

TABLE 2

Means and Standard Deviations of Variables[1]
N = 46,687 to 55,155

		Mean	Standard Deviation
Scale A:	Course Org. & Planning	4.27	.46
Scale B:	Communication	4.31	.44
Scale C:	Faculty/Student Interaction	4.31	.48
Scale D:	Assignments, Exams, & 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[2]	2.70	.35
Item 40:	Overall Evaluation	4.02	.51
Item 41:	College-Required vs. Choice[3]	.63	.32
Class Level[4]		.32	.46
Class Size[5]		.32	.47
Institutional Type[6]		.68	.47
Expected Grade[7]		4.53	1.25
Teaching Method: Lecture/Discussion[8]		.59	.49
Teaching Method: Discussion/Lab[9]		.32	.47

[1] Scales A through F and item 40, Overall Evaluation, were dependent variables; all others were independent variables, 5=high, 1=low.
[2] Mean of three items with 1 = difficult, fast, 3 = about right, 5 = elementary, slow.
[3] 63% of classes in students' major, or minor or as electives.
[4] 32% of classes at freshman/sophomore level; 68% jr./sr.
[5] 1 = 6 - 15, 0 = 16 or larger.
[6] 68% were four-year colleges/universities; 32% were two-year colleges.
[7] 1 = below C, 7 = A, 4 = B, 5 = B+
[8] 59% classified by instructors as primarily lecture/discussion classes.
[9] 32% classified by instructors as primarily discussions, labs, or labs with lectures.
9% classified by instructions as primarily lecture.

TABLE 3

Correlation Matrix
N = 46,687 to 55,155

	A	B	C	D	F	G	H	40	41	Cl Siz	In Typ	Cl Lev	Ex Gr	Lec
Scales[1] 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. Siz	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

---

[1] 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 = 2465		Business N = 5446		Education N = 3693		Social Science N = 9787		Fine Arts N = 3171		Natural Science N = 10,590		Eng. & Tech. N = 6397		Humanities N = 12,943	
Scales	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 = 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[1]

	Health N=2465		Business N=5646		Education N=3693		Social Science N=9787		Fine Arts N=3171		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.
Student Effort Inv. Scale G	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
Crse. Diff., Work, Pace Scale H	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
Course Outcomes Scale F	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. & 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

[1] See Table 2 for explanation of responses.

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 to 55,155
Standardized Beta Weights[1]

Controlled Variables	Overall Eval. N=53,549	A Crse. Org. & Planning N=53,515	B Communi-cation N=53,388	C Fac./Stu. Interaction N=53,297	D Assign., Ex., Grad'g N5=3,297
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
<u>Other Variables</u>					
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
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-Square	.72	.64	.67	.56	.68

[1] All t values were significant at .01 level unless indicated by ns or excluded.