Grading The Teacher: Reassessment Of Faculty Assessment

Sharon Tkacz
Kent State University—Geauga

Abstract

Faculty ratings were investigated along with variables such as expected grade, GPA, and class size. "Easy Graders" were identified quantitatively as instructors who gave higher average course grades than would be expected on the basis of the average GPA for students in a course. The grading leniency hypothesis was supported: easy graders received higher ratings. Comparisons between a branch campus and two departments at the parent campus of the same university demonstrated that ratings of branch faculty are higher in part because branch faculty are easier graders. These data show that student ratings reflect student opinions or liking of an instructor and may or may not reflect quality teaching.

Introduction

Student ratings of faculty are used in most colleges and universities to evaluate instructional effectiveness. In a review of research on the validity of student ratings and effective instruction, Abrami, Apollonia, and Cohen (1990) suggest two fundamentally different views of validity. One view is that student ratings are valid if they reflect what students learn, i.e., the quality of instruction. The other view is that student ratings are illustrations of students' opinions, regardless of whether they reflect what students learn.

McKeachie (1997) notes that what "teaching effectiveness" means must vary to some extent because different instructors will have different course goals across different contexts, such as traditional lecture, seminars, computer–assisted instruction, lab courses, or distance learning. Tasks might include memorization, knowledge acquisition, procedural skill learning, development of critical thinking, or simply an appreciation of

the discipline. In contrast, students may define effective teaching in terms of how easily learning occurs, so that an instructor who requires only passive listening and gives multiple choice tests is more effective than one who demands active involvement, gives essay exams, and requires out–of–class projects. Without an adequate definition of effectiveness, developing a valid instrument to assess it is impossible.

Other research has demonstrated that instructor ratings are influenced by obviously superficial and inappropriate variables. Morris, Gorham, Cohen, and Huffman (1996) found that perceptions of instructors vary as a function of fashion (i.e., formal or casual attire) and gender. They showed that clothing that enhances judgements of sociability and empathy *decreased* judgements of being well–informed and interesting and that both the gender of the student and the gender of the instructor influenced ratings. Yanowitz and Yanowitz (1997) found that both female and male students gave higher evaluations to male professors than female professors. They concluded that common stereotypes of men as being more competent and knowledgeable in their field of expertise contributes to these findings.

In an experimental investigation of instructor enthusiasm, Williams and Ceci (1997) manipulated presentation style of an instructor and found that both instructor and course ratings improved significantly. They compared fall and spring sections of a course taught by an experienced instructor who had recently attended a "teaching skills" workshop where he learned an enthusiastic presentation style (in terms of voice pitch and hand gestures). The *identical* course was delivered both semesters, using identical syllabi, texts, quizzes, exams, overheads, office hours, etc. However, in the spring, with the addition of "enthusiastic" style, ratings showed the instructor to be more knowledgeable, organized, accessible, and tolerant. Even course variables improved, with students reporting that during the enthusiastic presentation, the (identical) text was better, the (identical) grading policy was more fair, (identical) expectations and goals were more clearly stated, and that they learned more (even though point totals at the end of the semesters were nearly identical).

Given that there is disagreement over what teaching effectiveness is and that insignificant changes in an instructor's image can significantly alter ratings by students, it is reasonable to question the validity and

fairness of instructor evaluations. Recently, in the context of outcomes assessment, grade inflation has become a concern of colleges and universities. It seems reasonable to suggest that the increasing concern about grade inflation and the increasing reliance on instructor evaluations of faculty are related. One logical explanation is that faculty seek to improve their student evaluations by offering easier courses and assigning higher grades (i.e., a grading leniency hypothesis).

This investigation attempts to identify variables that (ideally) should be independent of faculty evaluations but which are, in fact, significantly related. Specifically, this study examines expected grade and class size, referred to by previous investigators as explanatory characteristics (Abrami et al., 1990) or background characteristics (Marsh, 1984) that significantly influence student ratings. It was hypothesized that students would give higher ratings to instructors from whom they expected higher grades. Further, it was expected that students would be fairly accurate in predicting their final grades. In addition, to support the grading leniency hypothesis, data were examined to determine whether higher grades were assigned to students with higher cumulative GPAs, or whether higher grades were assigned by those faculty who receive higher ratings (i.e., easy graders).

Method

Computerized summaries of instructor ratings were collected over three academic years (fall of 1994 to fall of 1997) and included over 17,000 students in 510 class sections. Only sections taught by full–time faculty (tenured, tenure–track, and non–tenure–track) were included in the study in an attempt to create a relatively homogeneous pool of instructors. The data from the parent campus of a state university included 93 sections taught by thirty faculty in the psychology and 181 sections taught by eighteen faculty in the history department. Regional campus data from one regional campus of the same university included 236 sections from all eight full–time faculty in six disciplines: English, computer technology, mathematics, psychology, accounting/economics, and business management. Regional campus sections included 46.3 % of the total; parent campus sections (history and psychology) included 53.7%.

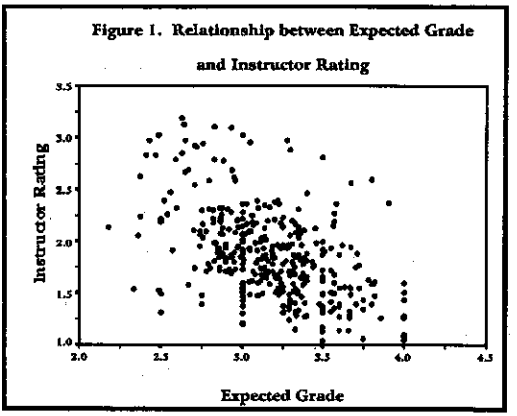
Variables consisted of instructor ratings by students, course characteristics, and student performance measures. Instructor rating indicated the students' opinion on a scale from 1 (excellent) to 6 (very poor), taken from one summary item ("overall, the instructor's teaching was") of the standardized KSU Instructor Report. Class size was the number of students enrolled in each section.

Student performance measures included expected grade (students' prediction of the grade they would receive), course grade (actual grade earned), and student GPA (cumulative GPA). These were calculated for each section (on the A = 4.0 scale). Course grade and GPA were obtained from the Office of Information Services, which provided grade distribution reports for each department after grades were handed in.

Results

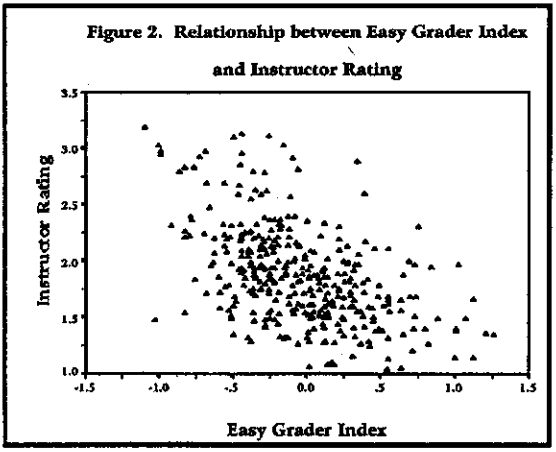
Both correlational and inferential analyses were performed. First, a multiple regression analysis determined which variables influenced instructor rating. In addition, regional campus and parent campus data were compared to test the hypothesis that more grade inflation occurs at regional campuses relative to the parent campus.

The most important variable affecting instructor rating was the expected grade in the course. Overwhelmingly, faculty who are perceived as "easy graders" received excellent ratings, indicated by a significant correlation between expected grade and instructor rating, $r = -.504$, $p < .001$, as shown in Figure 1. (Recall that the rating scale assigns "1" to excellent instructors and "6" to very poor instructors.) This correlation



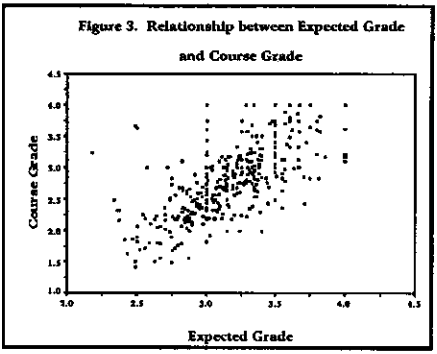
indicates that there is a relationship between the grade students expect and the ratings they give faculty: the higher the expected grade, the better the rating is for the instructor. This value is close to correlations reported by Williams and Ceci (1997) and Greenwald and Gillmore (1997): 0.42 and 0.45, respectively.

To determine if instructors who actually are easy graders received higher ratings, an easy grader index was calculated for each course. The easy grader index was defined as the average course grade (mean of all assigned grades for each section) minus the average GPA (mean of the cumulative GPA for students enrolled in that section). This calculation provided a quantitative measure of how easy an instructor was in terms of how close the average grade for an individual course was to an independent measure of academic performance (GPA). An easy grader index greater than zero indicates that the average grade assigned by the instructor is higher than the average GPA; a negative easy grader index indicates that the average assigned grade is lower than the average GPA. The extent to which an instructor actually is an easy grader was correlated



with the instructor rating of $r = -.505$, $p < .001$, as shown in **Figure 2**.

Students were accurately estimating their future grades, as shown (in **Figure 3**) by a significant correlation between expected grade and course grade, $r = +.681$, $p < .001$. This result directly addresses the question raised by Johnson and Christian (1990) that students may not



know their final grades when evaluations are completed. While they may not *know* with certainty, they do predict accurately.

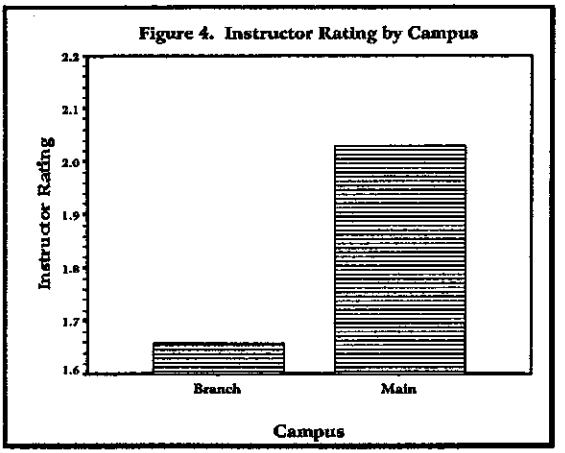
Stepwise multiple regression analyses, predicting the instructor rating from all of the variables mentioned above, support these interpretations, with the expected grade followed by the easy grader as the best predictors of instructor rating ($r = .675$, $p < .001$). Class size was also significantly related to these variables. The expected grade is lower in larger classes ($r = -.286$, $p < .001$), and students actually received lower course grades in larger classes ($r = -.409$, $p < .001$).

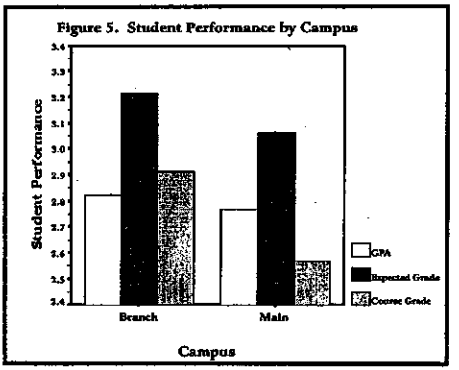
Campus Comparisons

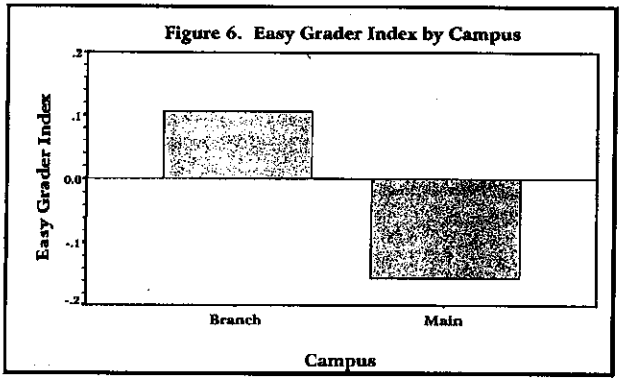
A multivariate analysis of variance demonstrated overall differences between the branch and main campuses ($F = 20.3$, $p < .001$). As shown in **Table 1**, all univariate means are reliably different. Regional campus students rate their instructors higher than parent campus students (**Figure 4**). McKeachie (1997) has suggested that ratings from regional campus students (who take predominantly first– and second–year courses) may be less valid because these students are less experienced and sophisticated educationally. In this sense, branch students may also be easy graders compared to more advanced students in advanced courses at the main campus. In addition, actual course grades and expected grades were higher at the branch campus compared to the main campus (**Figure 5**).

Finally, instructors at the branch campus actually are easier graders (**Figure 6**). In fact, if an easy grader index of zero indicates that an instructor is "fair" in the sense that course grades match GPA, parent

campus faculty are "hard graders" and on the average assign course grades







lower than a student's GPA.

Discussion

These data show that a number of variables influence ratings of teaching effectiveness. Most important is expected grade, students' prediction of their grade, followed by the easy grader index. These relationships between evaluations of instructors and evaluations of students suggest that students are not being objective. Further, their evaluations of instructors are not valid *in terms of teaching effectiveness.* Instead, evaluations reflect *satisfaction* with the instructor and/or the course.

The grading leniency hypothesis was supported: easy graders received higher ratings. Student ratings reflect student subjective opinions about an instructor. In addition, there is evidence that grade inflation results in both higher course grades for regional campus students and higher teaching ratings for regional campus faculty.

Class size was also consistently related to ratings, such that instructors with larger class enrollments per section received poorer ratings. There is previous research indicating that teachers do better in small classes (McKeachie, 1997). A reasonable interpretation is that increased interaction and individual attention are facilitated in small classes. However, this view also suggests that faculty who contribute most (in terms of FTEs) will essentially be penalized. In other words, those with the largest sections (e. g., several hundred students) are generating more income for a university than those with small sections, but their inability to have personal contact with their students may inequitably result in less favorable ratings.

Trout (1997) provides evidence that students want, "indeed, *demand* a dumbed–down educational experience." His analysis of narrative rating forms indicate that "students simply *do not want the rigorous and informative instruction the rating systems presume*" (p. 25). The question of fairness must be raised when invalid and subjective judgements of instructor effectiveness are used for important decisions such as tenure, promotion, and merit increases. Greenwald and Gillmore (1997) claim that the relationship between ratings and expected course grade is "well established." If true, grading leniency results in a bias such that "the goals of pedagogy and high instructor evaluations are in direct opposition" (p. 1209).

Recommendations

A number of recommendations have been made that try to resolve these issues. First, student ratings should provide only one component of a more comprehensive system of faculty evaluation. They alone should not comprise the only method of faculty assessment. Peer evaluation could provide another component.

Greenwald and Gillmore (1997) have suggested a statistical correction be used to remove the undesirable influences of variables such as grading leniency, course level, class size, etc. They believe that ratings can provide useful information about how well a course is liked or whether a student will enroll in another course by the same instructor. Finally, McKeachie (1997) has recognized a "lack of sophistication of personnel committees who use the ratings" (p. 1218) as the problem. He argues that comparisons *between* teachers are fundamentally unfair, particularly when comparing different courses and different disciplines. The use of numerical ratings to determine norms also has the distinct disadvantage of forcing 50% of faculty to be considered "below average." Rather than trying to interpret non–significant decimal point differences between ratings, he suggests that we train personnel committees to make better, more valid use of evaluation data. To correctly understand these statistical data, a background in statistics and a familiarity with the research on the numerous variables that unfairly influence instrutor ratings seems necessary.

A classroom demonstration of "relative deprivation used by Singleton (1978) shows how the current faculty assessment methods can be expected to lead to further grade inflation. Relative deprivation occurs when students compare themselves and the value of their grades with others and *perceive* they are less well off. They feel successful when their current achievements rise above previous accomplishments. They will be more satisfied with their grade the higher it is *relative to others*. Consequently, as grade inflation occurs, what used to be considered a "good grade" no longer is. Relatively speaking, they feel deprived and rate their instructors accordingly.

There is obviously a danger that dumbing down courses is the easiest way to improve one's student ratings. The implications for higher education should concern us all. According to Trout (1997), "The most effective device for lowering standards is the numerical evaluation form. . . . It is hard to imagine a practice more harmful to higher education. . . . These forms are not just invalid and unreliable; they are pernicious" (p. 30).

References

Abrami, P.C., d'Apollonia, S., and Cohen, P. A. (1990). "Validity Of Student Ratings Of Instruction: What We Know And What We Do Not." *Journal of Educational Psychology*, 82, No. 2, 219–231.

Greenwald, A.G. and Gillmore, G.M. (1997). "Grading Leniency Is A Removable Contaminant Of Student Ratings." *American Psychologist*, 52 (11), 1209–1217.

Johnson, R.L., and Christian, V.K. (1990). "Relation Of Perceived Learning And Expected Grade To Rated Effectiveness Of Teaching. *Perceptual and Motor Skills*, 70, 479–482.

Marsh, H.W. (1984). "Students' Evaluations Of University Teaching: Dimensionality, Reliability, Validity, Potential Biases, And Utility." *Journal of Educational Psychology*, 76, No. 5, 707–754.

McKeachie, W.J. (1997). "Student Ratings: The Validity Of Use." *American Psychologist*, 52 (11), 1218–1225.

Morris, T.L., Gorham, J., Cohen, S.H., and Huffman, D. (1996). "Fashion In The Classroom: Effects Of Attire On Student Perceptions Of Instructors In College Classes." *Communication Education*, 45, 135–148.

Singleton, R. (1978). "Classroom Demonstrations Of Social Psychology Principles." *Teaching Sociology*, 5, 187–200.

Trout, P.A. (1997). "What The Numbers Mean: Providing A Context For Numerical Student Evaluations Of Courses." *Change*, 29 (5), 24–30.

Williams, W.M. and Ceci, S.J. (1997). "'How Am I Doing?': Problems With Students Ratings Of Instructors And Courses." *Change*, 29 (5), 13–23.

Yanowitz, K.L. and Yanowitz, J.L. (1997). "Students Like Male Professors: Attitudes As A Function Of Gender." Paper presented at the 9th annual American Psychological Society Conference, Washington, D. C.

Biography

Sharon Tkacz is the self–proclaimed "Queen of Psychology" at the Kent State University---Geauga. Her expertise is in spatial cognition. During her sabbatical this year, she is studying individual differences in how people read maps and process cardinal directions. Her Highness may be reached by e–mail at stkacz@geauga.kent.edu.