STUDENT EVALUATION OF INSTRUCTION: THE THEORETICAL EFFECTS OF LENIENCY AND RECIPROCITY

Dennis E. Clayson, University of Northern Iowa, College of Business, Cedar Falls, Iowa 50614-0126; (319) 273-6015

ABSTRACT

This paper looks at the controversy generated by a possible grade/evaluation association found in the student evaluation of teaching. It develops a theoretical framework to understand the contributions to the effect by both leniency and reciprocity, and shows that the two are confounded when looking only at leniency. The findings suggest that reciprocity is the appropriate starting point for a study of the association.

INTRODUCTION

The student evaluation of teaching (SET) has come under close scrutiny. One finding that continues to reappear is an association between grades that an instructor gives and the evaluations students give to the instructor. This relationship has been studied closely because of its implication for the validity of the SET process. The present study expands this discussion of the association by looking at the theoretical relationship between two proposed effects.

What is currently known?

Instructors and students believe that there is a grades/evaluation association. A vigorous debate has developed about the accuracy of this belief. Early studies found the relationship (Feldman 1976, and Stumpf and Freedman 1979), but it was explained as a statistical artifact (Seiver 1983). More recent research has shown that the association may exist outside of a simple least squared (SLS) relationship (Braskamp and Ory 1994; Marsh and Dunkin 1992), but a true grade/evaluation relationship was still denied (Cashin 1995; Marsh and Dunkin 1992; Kaplan, Mets and Cook 2000; Marsh and Roche 2000). Marsh and Roche (1999) referred to the idea that lower grades would result in lower student teacher evaluations as only a "presumption." Other research, however, claims to have found a relationship (Bacon and Novotny 2002; Bharadwaj, Futrell, and Kantak 1993;

Gillmore and Greenwald 1999; Goldberg and Callahan 1991). Marsh, Hau, Chung, and Siu (1997) found a significant difference between the grades students indicated they received from those instructors chosen as "good" and "poor" teachers. Wilhelm (2004) compared course evaluations, course worth, grading leniency, and course workload as factors of business students choosing classes. A conjoint analysis showed that, "... students are 10 times more likely to choose a course with a lenient grader, all else being equal" (p. 24). Studies, mostly from colleges of education, continue to claim that the grade/evaluation effect does not exist independent of confounding variables (see Marsh and Roche 2000 for a review).

Grade/Evaluation Hypotheses

Several hypotheses have been proposed to explain the apparent grade/evaluation relationship.

1. Leniency: The essence of the leniency hypothesis is "... that students will reward teachers who grade leniently with higher teacher and course evaluations" (Bacon and Novotny 2002, p. 5). It is important to this hypothesis to recognize that, "... it is not the grades per se that influence SETs, but the leniency with which grades are assigned" (March and Roche 2000, p. 204).

2. *Reciprocity:* Students reward the instructor who gives them good grades and withhold higher evaluations from an instructor who gives them a lower grade. This hypothesis states that students have a tendency to modify evaluations based on the grades they individually receive. It states that a student given an A will generally give an instructor a higher evaluation than a student receiving a C, irrespective of the general leniency of the instructor. It does not deny that the class effect may also be in play.

3. *Other hypotheses*: It has also be proposed that the relationship could be due to prior characteristics like the rigor of the instructor's grading policies, class workloads, motivation, and prior student interest in the class. Another hypothesis simply assumes that both the grades and the evaluations measure teaching effectiveness, and therefore reflect a valid association. Attribution has also been advance as a cause of the associations. Since learning and achievement are difficult for students to evaluate, they may infer the ability of the instructor to teach from the grade they receive.

Greenwald and Gillmore (1997) found evidence that contradicted all these hypotheses except the leniency explanation. Marsh and Roche (2000) countered most of their claims and indicated that they found little support for the leniency hypothesis. Since the reciprocity is a new concept, it was not considered by either set of researchers.

PROBLEMS

Greenwald and Gillmore (1997) did not differentiate clearly between leniency and reciprocity. Marsh and Roche's (2000) work is clearly aimed at leniency. The literature review of both papers indicates that leniency and reciprocity are either confounded by methodology, or that reciprocity was not considered. Part of this problem is methodological. As Marsh and Roche point out, the appropriate case for a study of leniency is a class. The appropriate case to study reciprocity is the student. These are statistically distinct concepts.

PURPOSE OF STUDY

This study looks at the hypothetical relationship between leniency and reciprocity in the grade/evaluation effect. Outside of education colleges, researchers claim that leniency exist, but only two sources (Clayson 2004; Stumpf and Freedman 1979) have identified the possible confounding effects of reciprocity. The study sets the foundation for future research by looking at the theoretical relationship that could exist between the two effects and some of the implications of these interactions.

Development

It is possible for leniency and reciprocity effects to exist independently or in conjuncture with each other. It is also possible for a leniency effect to appear to exist falsely. Consequently, it is important when looking at these combinations to distinguish between theoretical and observed conditions. The theoretical leniency relationship is given by:

$$Eval_m = \beta_{0j} + \beta_{1j} Grade_m + e_j .$$

The theoretical reciprocity relationship is:

$$Eval_{ij} = \beta_{0ij} + \beta_{1ij} Grade_{ij} + e_{ij} .$$

It may appear as if a leniency effect exists even when it does not; consequently the observed value is summarized as:

$$Eval_m = B_{0j} + B_{1j} Grade_m + e_j .$$

The relationship between leniency and reciprocity effects can be seen in six possible combinations.

Condition 1: Both leniency and reciprocity exist. The theoretical pure leniency relationship is assumed to be:

$$Eval_m = \beta_{0j} + \beta_{1j} Grade_m + e_j .$$

Within a class each student is also reacting as:

$$Eval_{ij} = \beta_{0ij} + \beta_{1ij} Grade_{IJ} + e_{ij} .$$

It may be perceptually easier to deal with covariance in terms of the much-used correlation coefficient, hence the observed leniency effect would include both conditions so that the apparent association would be:

$$r_{1j} > \rho_{1ij} \text{ and } r_{1j} > \rho_{1j} .$$

Condition 2: Leniency does not exist, but reciprocity does. The theoretical relationship is now: $\beta_{1j} = 0 : Eval_{ij} = \beta_{0ij} + \beta_{1ij} Grade + e_{ij}$. Under these conditions $Cov(Eval_m, Grade_m) < Cov(Eval_{ij}, Grade_{ij})$, consequently the observed leniency effect would be that $r_{1j} >> \rho_{1ij}$.

Condition 3: Leniency exists, but reciprocity does not exist. The theoretical relationship is now $Eval_m = \beta_{0j} + \beta_{1j} Grade_m + e_j$ while $\beta_{1ij} = 0$. If $e_j \sim N(0, \sigma^2)$, then $r_{1j} = \rho_{1j}$ with large n's. The problem with this condition is not mathematical, but logical. The condition could only be true if each student responded to the overall grading average (or perhaps to some real or imagined standard) of the class, but not to their own grade. How students could obtain such a group mind is conceptually difficult, especially if their own grade deviated significantly from the norm. It would be interesting to see if instructors who publish a grade average, which they attempt to confirm with grading practices, would demonstrate more or less of a leniency effect.

Condition 4: Neither leniency or reciprocity exist. In this case, both $\beta_{1j} = 0$ and $\beta_{1ij} = 0$. There should be no observed leniency effect with large samples.

Condition 5: Leniency exists, but reciprocity is either a function of other variables or only one of many variables that influences the evaluations. A simplified example would be: $Eval_{ij} = \beta_{0ij} + \beta_{1ij} Grade + \beta_{2ij} x_1 + ... \beta_{nij} x_{n-1} + e_{ij}$ where x_1 through x_{n-1} are student, instructor, or class variables that could also modify the evaluations. It is also possible that both $Eval_{ij}$ and $Grade_{ij} = f(x_1, ... x_{n-1})$. The observed value of r_{1ij} from

$$Eval_{ij} = B_{0ij} + B_{1ij}Grade + e_{ij},$$

without taking the other variables into account, could be almost anything. Without proof, the measured r_{1j} could be either increased or decreased by the influence of r_{1ij}.

Condition 6: Leniency does not exist, but reciprocity is either a function of other variables or only one of many variables that influences the evaluations. Under this condition, the observed leniency effect (r_{1j}) would simply be a function of the apparent r_{1ij}, which in turn is being modified by variables x_1 to x_{n-1}. Depending upon the selection of classes, almost any outcome is possible. For example, if a significant, but unknown *x* variable was the age of the student, it is possible that a sample of classes from a business school, in which the average student is older, may find an apparent positive r_{1j}, while classes selected from educational colleges, which may have a lower average student age, could have an average $r_{1j} = 0$. In this case, it would appear to a researcher from educational colleges that leniency did not exist, and a business researcher would conclude the leniency did exist. In both cases only the age effect actually existed.

Simulation

To test these conditions, a large random sample of classes would be needed. Consequently, it is difficult to investigate the predictions of Conditions 1 through 6. There have been several such studies, but they have been interpreted in different ways. Johnson found both leniency and reciprocity as did Stumpf and Freedman, but Johnson did not consider reciprocity and Stumpf and Freedman did not identify it as such. Marsh and Roche found a grade/evaluation effect but denied that it was a leniency effect. Therefore, a simulation was created to give some guidance on what would be expected under each condition. Each run simulated the effect of a given condition on 20,000 classes, with 30 students in each class.

Condition 1 (*Leniency yes, Reciprocity yes*): The program assumed that Evaluation Means = *f*(Grade Means). A student grade (n = 600,000) was selected at random with a given class leniency, target mean grade (M) that could range from 0.5 to 3.5, and such that grades ~ N(M, 0.68). The standard deviation of 0.68 was calculated from actual grades from a sample of 600 business students at the writer's university. The student evaluation in the reciprocity condition was Eval = b(grade) + a, where b was calculated from a correlation r ~ N(0.21, .07). The leniency class effect was calculated as: $Eval_j$ = b_j(class mean) + a_j, where b_j was derived from r_j ~ N(0.31, 10). Both the mean of r and r_j are from the large study at Duke (Johnson 2003). It is not known how reciprocity and leniency interact. The simulation simply assumed, therefore, that each made an equal contribution so that $Eval_{if}$ = .5 (Eval) + .5($Eval_j$). The initial simulation did not consider other possible influences. At almost all leniency targets, theapparent leniency correlation was larger than the starting values of r and r_j.

Condition 2 (*Leniency no, Reciprocity yes*): The program was the same as that above except there was no contribution made by leniency. As predicted, a very strong apparent leniency effect was found, although none actually existed.

Condition 3 (*Leniency yes, Reciprocity no*): The program was the same as Condition 1 except that no reciprocity effect was inputted. Leniency appears to exist and is roughly equal to the inputted leniency association. It also appears to decrease as the leniency of the instructor increases.

Condition 4 (*Leniency no, Reciprocity no*): All levels are randomly assigned. As would be expected, all associations are essentially zero.

Condition 5 (*Leniency yes, Reciprocity mixed*): The program was essentially the same except that a second variable called X_2 was added on the class level. As an illustration, X_2 was selected as an actual variable from an existing data set at the writer's university. It had r = .07 with evaluations and r = .73 with grades. The apparent leniency effect was greatly reduced from the pure Condition 1 example, and follows a different pattern. It is interesting in this condition to run simulations of a small number of classes, in this case ten (300 students). Seven average leniency target grades were selected (ranging from 0.5 to 3.5 on a scale from 0 to 4), and ten simulations of each was run (70 total). The apparent leniency correlation ranged from -0.480 to 0.815. Although leniency was actually present, 84% of the samples showed no significant ($p < .05$) leniency effect.

Condition 6 (*Leniency no, Reciprocity mix*): The program was the same as Condition 5, but leniency was removed. There still appears to be a leniency effect (although none exists). In mid-ranges of leniency targets, this condition could not be differentiated from Condition 5. Again simulations of a small number of classes (10 classes) were run. The apparent leniency correlation ranged from -0.417 to 0.821. Although leniency was not present, 17% of the samples showed a significant leniency effect.

DISCUSSION

The apparent leniency effect can be seen as resulting from actual leniency, from reciprocity, or from a combination of the two. Each condition suggests a slightly different pattern of apparent leniency results. For example, when leniency actually exists there seems to be a general negative relationship between the leniency target of the instructor and the apparent leniency effect. Reciprocity, when predominant, does not show this pattern. Leniency effects create an interesting prediction. The average grade in the writer's business college is 2.58, while a corresponding average in one department of the education college is 3.72. The simulation suggests that an apparent leniency effect would be shown to be stronger in a business college sample than in an education college, with no change whatsoever in actual student or instructor behavior between the two. For example in Condition 1, the simulation estimated that $r = 0.59$ in business and $r = 0.37$ in education. In Condition 3, the apparent leniency in business classes would be $r = 0.40$ and $r = 0.25$ in education classes.

Classrooms are obviously more complex than these initial simulations. The effects of other intervening variables would generally be to lower the associations. The analysis does demonstrate the complexity of the problem. Leniency can appear to exist falsely, and is severely confounded by reciprocity. The largest associations predicted and found in the study were from classes that did not have any contribution by leniency. It would appear that researchers have been attacking this problem from the wrong perspective. As long as the studies continue to emphasize leniency, the results will continue to be highly variable. From both methodological and theoretical perspectives, a firm foundation and understanding of reciprocity is needed before a leniency effect can be fully studied.

References available upon request from the author.