

The Relationship Between Students' Grades and Their Evaluation of Instructor Performance

Sherry D. Salmons
DePaul University

Instructors who receive higher ratings from their students are often believed to be superior in their teaching ability, and as a result, often receive substantial rewards (i.e. raises, grants, tenure) based largely on these ratings. Because of the importance of student ratings, this research examined the impact of students' grade expectations on their evaluation of instructor performance. Four hundred and forty-four Radford University students served as voluntary participants and completed two separate instructor evaluations during the semester: A preliminary student evaluation of the instructor was completed before the first class exam was given to eliminate grade contamination and a second instructor evaluation was taken at the end of the semester, which included a question concerning expected grade. Results indicated that students who expected to receive a "F" significantly lowered their evaluation of the instructor, whereas, students who expected to receive an "A" or "B" significantly raised their evaluations. These findings support the idea that instructors receive higher ratings in part because they are more lenient in their grade assignment. However, student grades only accounted for a small percentage of the evaluation variance.

Much emphasis is placed on the evaluation of classroom instruction as often those instructors receiving higher ratings from their students are perceived to be superior or outstanding in their ability. As a result of these perceptions, Kohlan (1973) suggested that administrative decisions regarding promotions, salary, and tenure are increasingly including student evaluations as a substantial factor. If these student evaluations are an accurate assessment of ability, then this reward process appears equi-

table. However, it has been argued that student ratings merely reflect the likability of professors rather than their ability to instruct. Moreover, Ducette and Kenney (1982) report that the more lenient an individual's grading policy, the more likable they become; particularly for courses required by the university.

To determine the relationship between student grades and ratings of instructor performance, four different research designs have typically been used: 1) Correlational studies of grades and ratings, 2) Complex correlational studies, 3) Experimental manipulation of grades, and 4) Correlational studies on consistency and stability of early and late semester evaluations.

Correlation of Grades and Ratings

In reviewing the literature on student ratings, it is clear that the results regarding the relationship between grades and instructor evaluations are very inconsistent. Though some research has reported a positive correlation between expected grades and student evaluations (Centra & Linn, 1973; Hildenbrand, Wilson, & Dienst 1971; Granzin & Painter, 1973), other research has found no such relationship (Blum, 1936; Ducette & Kenney, 1982; Holmes, 1972; Voeks & French, 1960).

Additionally, path analytic studies have found a small, yet significant relationship between grades and evaluations but suggest that other factors such as motivation and satisfaction also determine the ratings (Howard & Maxwell, 1980; 1982). Though Howard and Maxwell do not discount the effects that expected grades have on student ratings, they suggest the following relationship as a natural progression: (a) good teaching leads to better learning, (b) greater student motivation leads to better learning, (c) greater student learning results in higher grades, and (d) greater student learning leads to greater student satisfaction (Howard & Maxwell, 1980). The results of Howard and Maxwell (1980) indicate a much larger relationship exists between students' motivation and their satisfaction with an instructor ($r = .48$) than exists between expected grade and satisfaction ($r = .23$).

Stumpf and Freedman (1979) further suggest that a large amount of variance may be confounded by the instruments used, the institutions studied, and the statistical analysis utilized. Analyses have ranged from nonparametric, such as sign tests and chi-square, to univariate and multivariate analysis of variance (Brandenburg & Slinde, 1977). The variation

of questionnaires utilized, as well as the statistics used to interpret the data, could account for the inconsistent results regarding the relationship between student grades and ratings of instructor performance.

Anikeef (1953) reported that the extent to which students evaluate faculty according to the grades they receive varies among the academic levels of students. For example, she found no significant relationship between grades and ratings for junior and senior level courses, yet found a significant relationship for freshmen and sophomore level courses. This difference could be an indication of immaturity which results in cognitive dissonance or possibly a reflection of the fact that most freshmen and sophomore courses are required by the university rather than elected by the student.

The results of Ducette and Kenney (1982) show a strong relationship between expected grade and ratings in graduate core and statistics courses. Ducette and Kenney hypothesize that because students are required to take these courses rather than being given a choice, grades become a larger determinant of instructor ratings (Ducette & Kenney, 1982). Still, they argue that the grade a student receives in a course, to some extent, does reflect how much the student has learned. Though there is research which indicates a significant relationship between expected grades and instructor evaluations (Granzin & Painter, 1973; Miller, 1972; Pratt & Pratt, 1976), the evidence is less clear on the relationship between actual final grades and evaluations (Endo & Della-Piana, 1976; Frey, 1976; Granzin & Painter, 1973; Treffinger & Feldhusen, 1970). These findings indicate that students are partially influenced by the grade they expect to receive and this expectation may or may not reflect their actual final grade. Apparently the students are basing their rating on the grade they are expecting to receive and that expectation is often inaccurate.

Ostensibly, inconsistencies exist within the research results. However those inconsistencies dissipate when the design being used and the variable being studied are considered. For example, Ducette and Kenney (1982) looked at the relationship between each individual student's grade and the rating he/she assigned to an instructor and found a correlation of .10. Whereas, Frey (1976) considered the relationship between mean class grades and mean instructor ratings and found a correlation of .75. The results seem to indicate that individual students are less influenced by expected grade than has been reported. Moreover, the significant relationship appears to lie within the overall class grades. In other words,

because the relationship between an individual's rating and their expected grade is smaller than that of the class ratings and expected grades, it can be argued that there is less rating influence on an individual basis. Howard and Maxwell (1980; 1982) have explained this phenomena in their path analytic research and point to better teaching ultimately leading to higher ratings. An instructor with high class grades and high ratings simply could be a better instructor.

Moreover, it is necessary to consider whether the evaluation is based upon the grade the student expects to receive or one the student actually receives. If the correlation between a student's expected grade and evaluation is higher than the correlation of the actual grade received, this would indicate students are partially basing their evaluations on expectation of grade. On the other hand, if the strength of the correlations were reversed, this would indicate students are not necessarily persuaded by the grade they expect to receive.

A meta-analytic review of the literature using the standard procedures outlined by Hunter and Schmidt (1990) was conducted for this thesis by first separating studies based on the four factors listed above. As shown in Table 1, the results of this analysis indicate that the type of research design used has a definite effect on the relationship between student ratings and grades. The effect becomes apparent by the fact that the correlation between individual student's expected grades and their instructor ratings is higher than that of the correlation between the individual student's actual grades and their instructor ratings. Apparently individual students are partially basing their evaluations on the grade they expect to receive and some bias does exist for or against the instructors depending on the instructor's grading policy.

The same effect was seen when considering the relationship of the average expected grades in a class and the mean teacher ratings for a class versus the average of actual class grades and the mean teacher ratings. However, the overall correlation of the class ratings to grades was almost double the correlation of individual students ratings to grades. The unit of measurement (student versus class) is an obvious factor responsible for variance in the instructor ratings. In other words, although there is some rater bias on the student level it is not nearly as high as previously believed. If the class correlation is higher than the individual correlation, other factors are going into the instructor rating (i.e. effective teaching).

Complex Correlational Models

Granzin and Painter (1973) looked at a variety of factors which could possibly influence instructor ratings (i.e. ease of course, grade expected, grade capable of, grade deserved, interesting/entertaining, contribution to vocation, and contribution to general education). Of the factors considered, interesting/entertaining ranked as the highest ($r = .59$), with contribution to general education ranking second ($r = .46$). Students' expected grades ranked significantly lower with a correlation of only .16. The authors deduced that because the highest correlations fell into the interesting/entertaining category, jokes, theatrics, simple well-chosen materials, and well-delivered lectures are of major importance in receiving high instructor ratings (Granzin & Painter, 1973).

Marsh (1980) agrees that other factors are prevalent in instructor evaluations and considered what effect the following four variables had on instructor ratings: Prior Interest In The Course, Work Difficulty, General Interest In The Course, and Expected Grade In The Class. He found that of the four, Prior Interest and Expected Grade both account for only four percent of the variance and Work Difficulty ($r = .14$) and General Interest ($r = .12$) slightly less. Only 11% of the variance was accounted for by a combination of the four variables indicating other factors are more influential on the final ratings. As Marsh points out, if students' evaluations are biased, the bias is not a simple one.

Experimental Manipulation of Grades

Vasta and Sarmiento (1979) investigated, in a natural setting, the effects that liberal versus stringent grading might have on instructor evaluations. The authors team-taught two large undergraduate courses and kept all quizzes, assignments, lectures, and exams at a constant for both classes. The only variation in the classes was the distribution of grades either by being exceptionally liberal or stringent in their policy. One of the sections received 25% A's or B's and 37% D's or F's, while the other section received 48% A's or B's and 15% D's or F's. The students in both sections were asked to complete an instructor evaluation form near the end of the semester. The results indicated that of the items concerning the instructor, 24 of the 32 were rated more favorably by the group in the liberally graded section (Vasto & Sarmiento, 1979). Though the authors argue this clearly implies a relationship between the two variables, they

fail to consider contamination of students comparing grade distributions from the two sections. In other words, the significant difference in ratings could partially be accounted for by the students' knowledge of the vast differences in the distributions.

Table 1

Average correlations across studies between student grades and instructor evaluations

Unit of Measure	Grade		
	Expected	Actual	Total
Student	.24	.11	.23
Class	.42	.23	.29
Total	.28	.22	.26

Stability and Consistency of Ratings

A fourth way of looking at the relationship between grades and evaluations is to investigate the stability of ratings over the 16 week semester. The thinking behind this method is that a high correlation between the two ratings would indicate a consistent opinion of an instructor and that grade bias is not a significant factor in altering that opinion.

Kohlman (1973) conducted a pre and post evaluation on an individual student basis which was kept anonymous by coding. He administered an Instructor Evaluation Questionnaire (IEQ) to 271 undergraduate students the second day of class and then again during the last week of the semes-

ter. The results indicated a significant relationship between early and late IEQ scores on overall instructor ratings ($r = .58$) (Kohlan, 1973).

Tagiuri (1969) reported the stabilization of ratings as the phenomenon of primacy which is the tendency for data obtained at the beginning of the formation of an impression to remain as a salient feature, thereafter, unless strongly contradicted. Therefore, the overall impression given the first day of class is apparently a very important and lasting impression upon the students' ultimate appraisal of instruction. Though it may be concluded such early stabilization invalidates this method of collecting instructor evaluations, it can also be argued that early instructor behavior may possibly foreshadow teaching effectiveness and, thus, be highly related to foster student learning (Kohlan, 1973).

Statement of Purpose

Based on previous studies, it is apparent that further research is necessary to determine the relationship between students' expected grades and their subsequent evaluations of the instructor. By strictly using a correlational or complex correlational research design, the results do not indicate the cause of the relationship between grades and ratings. The third method reviewed (manipulation of grade distributions) is not only ethically questionable, but possibly invalid because the evaluations are contaminated by students from the two sections discussing the very different grading policy. Finally, by simply looking for rating stability (the fourth method in review), the extent to which expected grade alters that stability is not considered.

The present study attempted to control for grade contamination in ratings by evaluating teaching performance prior to the first exam being administered. Although it was only the third week of the semester, it was believed to be ample time for students to have formed impressions about their instructor. This rating was then compared to an evaluation taken later in one semester after students had received feedback (i.e. grades) on five of their six exams.

Hypothesis: The correlation between student's expected grade and the ratings of their instructor taken before the first test will not be significantly different from the correlation between expected grade and the instructor ratings taken after the students had received feedback on test performance.

METHOD

Subjects

Subjects were 444 (146 Male, 298 Female) college students from a public university with an enrollment of about 9,000. The subjects were introductory psychology students, ranging in age from 18 to 22, who participated in the study as part of their normal class procedure.

Procedure

Participants were asked to complete two different instructor evaluation forms on the graduate teaching fellows instructing their class. The first evaluation was taken three to four weeks into the semester (prior to the first exam being given) and the second was taken during the 13th week of school. The first evaluation form consisted of six questions asking about the student's race, major, gender, year in school, mother's maiden name, and an overall evaluation of their instructor's performance to date. The subjects were asked to write their mother's maiden name on the form so that their first evaluation could later be linked anonymously to their second evaluation. It was assumed that this initial rating was absent of grade contamination because the student had not yet had the opportunity to succeed or fail in the class.

The university's official evaluation form was used to gather the second rating. This evaluation form consisted of 15 questions ranging from the fairness of the grading policy to the number of times the instructor missed class. The form also included a question asking about the grade the student expected to receive in the course, as well as a question asking for an overall evaluation of the instructor. The subjects were again asked to write their mother's maiden name on the form.

RESULTS

To test the hypothesis, individual student ratings from both the first evaluation and the second evaluation were correlated with the grades that students expected to receive based on their test scores going into the final exam. The two correlation coefficients were then transformed into z

scores and the difference between the two coefficients was analyzed using Fisher's r to z formula (Ferguson, 1981).

Though grade expectancy did not significantly correlate with the first evaluation ($r = -.01$), it did correlate significantly with the second evaluation ($r = .24$, $p < .0001$). Moreover, there was a significant difference between these two correlation coefficients ($Z = 2.94$). Though these results do not support the hypothesis that there would be no difference between the two correlations, grade expectancy only accounted for a small percentage of the evaluation variance (approximately five percent).

To determine the practical significance of the above results the difference in correlations was transformed into a d score and an effect size of .26 was found. The effect size was then multiplied by the standard deviation of the ratings. Based on the results of the procedure, it appears that on the university's five-point rating scale, expected grade accounted for .23 of a rating point. While this amount is not insignificant, it is a rather small effect size and probably not considered a crucial amount when making personnel decisions. For example, if one considers an instructor who consistently receives a mean rating of 4.4 but is considered a liberal grader, it could be assumed that his/her rating would only drop to a 4.2 by becoming a strict grader (which is still a good rating). Moreover, an ineffective instructor whose mean rating is a 2.5 and who is considered a very strict grader, will still be rated poorly (2.7) if he/she becomes a lenient grader.

To analyze the data in a different way, students were grouped based on the grade they expected to receive (i.e. all A's, B's, etc.) and their early semester ratings were subtracted from their late semester ratings in order to determine whether or not their ratings significantly changed after receiving test scores. As shown in Table 2, the results of this ANOVA and subsequent Least Significant Difference (LSD) tests indicated that grade expectancy had a significant effect on instructor evaluation $F(4, 389) = 6.52$, $p < .0001$. An individual t -test was also utilized to test for means significantly differing from zero and it was determined that students expecting to receive a "D" or a "F" significantly lowered their ratings from the first to second rating while students who expected to receive an "A" or a "B" significantly raised their evaluations.

To analyze the effect size of a specific grade expectancy the F score (from the ANOVA) was converted to a d score. The converted score shows an effect size of .49 which indicates a variance of .43 of a

Table 2**Mean changes in instructor ratings as a result of expected grade**

Expected Grade	n	Mean	Standard Deviation
A	51	.39*	.11
B	160	.20*	.06
C	129	.03	.07
D	42	-.23*	.13
F	12	-.58*	.24

* mean change score differs significantly from zero

point on the university's five point rating scale. Unlike the ratings variance (.23), this is a rather large amount and could account for many of the personnel errors being made which are cited by Kohlan (1977). For example, the instructor receiving a mean rating of a 4.0 and considered a strict rater, could very possibly deserve a rating of 4.4 which would elevate that instructor into a category for personnel perques he/she did not receive before.

DISCUSSION

The present study attempted to take previous research further by incorporating parts of the research designs and by eliminating others. A correlational research design was incorporated to establish the relationship between expected grade and instructor ratings and at the same time a stabilization of the ratings was considered by conducting an early and late semester evaluation. Moreover, just as the experimental design attempt-

ed to keep all other variables constant, only the graduate teaching fellows were considered in the study. It is apparent that there is a significant relationship between the grades students expect to receive and their subsequent evaluation of instruction. However, that relationship is a small one; accounting for only about six percent of the variance in the overall evaluation. Still, when considering the effect size on the typical university's five point rating scale, expected grade accounts for .2 to .4 of a point depending on which type of statistical analysis is utilized. While this amount is not minimal, it can be argued that many other factors are responsible for the variance in ratings. An instructor's lecture style, ability, and willingness to help outside the classroom are just a few of the other variables attributable to a student's evaluation. Furthermore, even when expected grade is factored out of instructor ratings, an ineffective instructor will still receive low ratings and excellent instructors will still receive good ratings.

A suggestion for further research would be to have the instructors' lecture style viewed by neutral subjects. Both highly rated instructors as well as low rated instructors would be videotaped and the subjects would view the tapes and rate them accordingly. Because grade contamination is eliminated completely, a high correlation between the neutral subjects ratings and the instructors' actual student ratings would indicate the previous assessment is an accurate one.

While it is necessary to remember effective instructors are rated on more than their interesting lecture style, this factor is still responsible for a large part of instructor rating variance (Granzin & Painter, 1973).

REFERENCES

- Anikeef, A. M. (1953). Factors affecting student evaluations of college faculty members. *The Journal of Applied Psychology*, 37, 458-460.
- Blum, M. L. (1936). An investigation of the relation existing between students' grades and their ratings of the instructor's ability to teach. *Journal of Educational Psychology*, 27, 217-221.
- Brandenburg, D. C. and Slinde, J. A. (1977). Student ratings of instruction: Validity and normative interpretations. *Research in Higher Education*, 7, 67-78.

- Centra, J. A. and Linn, R. L. (1973). Student points of view in ratings of college instruction. *Research Bulletin*, RB-73-60. Princeton, N. J.: Education Testing Service.
- Ducette, J. and Kenney, J. (1982). Do grading standards affect student evaluations of teaching? Some new evidence on an old question. *Journal of Educational Psychology*, 63, 130-133.
- Endo, G. T. and Della-Piana, G. (1976). A validation study of course evaluation ratings. *Improving College and University Teaching*, 24, 84-86.
- Ferguson, G. A. (1981). *Statistical Analysis in Psychology and Education*. New York: McGraw-Hill.
- Frey, P. W. (1976). Validity of student instructional ratings. *Journal of Higher Education*, 47, 327-336.
- Granzin, K. L. and Painter, J. J. (1973). A new explanation for students' course evaluation tendencies. *American Educational Research Journal*, 10, 115-124.
- Hildenbrand, M., Wilson, R. C., and Dienst, E. R. (1971). *Evaluating University Teaching*. Berkeley: Center for Research and Development in Higher Learning.
- 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. and Maxwell, S. E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education*, 16, 175-188.
- Howard, G. S. and Maxwell, S. E. (1980). The correlation between grades and student satisfaction: A case of mistaken causation? *Journal of Educational Psychology*, 72, 810-820.
- Hunter, J. E. and Schmidt, F. L. (1990). *Methods of Meta-Analysis*. Newbury Park: SAGE.
- Kohlan, R. G. (1973). A comparison of faculty evaluations early and late in the course. *Journal of Higher Education*, 44, 5897-595.
- Marsh, H. W. (1980). The influence of student, course, and instructor characteristics in evaluations of university teaching. *American Educational Research Journal*, 17, 219-237.
- Miller, R. I. (1972). *Evaluating Faculty Performance*. San Francisco: Jossey Bass.
- Pratt, M., and Pratt, T. A. (1976). A study of student-teacher grading interaction process. *Improving College and University Teaching*, 24, 73-81.

- Stumpf, S. A. and Freedman, R. D. (1979). Expected grade covariation with student ratings of instruction: Individual versus class effects. *Educational Psychology, 71*, 293-302.
- Tagiuri, R. (1969). Person perception. *The Handbook of Social Psychology, 3*, 395-449.
- Treffinger, D. J. and Feldhusen, J. F. (1970). Predicting student ratings of instruction. *Proceedings of the 78th annual convention of the American Psychological Association, 5*, 621-622.
- Vasta, R. and Sarmiento, R. F. (1979). Liberal grading improves evaluations but not performance. *Educational Psychology, 71*, 207-211.
- Voeks, V. W. and French, G. M. (1960). Are student-ratings of teachers affected by grades? *Journal of Higher Education, 31*, 330-334.