DOCUMENT RESUME

ED 118 165 JC 760 089

AUTHOR Powell, Robert
TITLE Grading Style and Student Evaluation of Faculty.
INSTITUTION William Rainey Harper Coll., Palatine, Ill.
PUB DATE Apr 75
NOTE 57p.

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage
DESCRIPTORS Annotated Bibliographies; Correlation; Evaluation Criteria; *Grades (Scholastic); Junior Colleges; *Literature Reviews; *Post Secondary Education; Student Attitudes; *Student Teacher Relationship; *Teacher Evaluation
IDENTIFIERS William Rainey Harper College

ABSTRACT

This paper discusses the association between student grades and student ratings of faculty. The first section reviews a 1974 study of Harper College English teacher ratings, which showed a correlation of .73 between the grades the teachers gave students and the ratings students gave the teachers. The second section reports the findings of a 1975 replication study which showed grade-rating correlations of up to .79. The third section provides a review of the literature in the form of an annotated bibliography, indicating that the Harper findings are typical of the findings of prior research at other colleges. Twenty-eight studies involving more than 70,000 student ratings of faculty in more than 50 colleges and universities have been conducted and published since 1954. In every study, at least some association has been found between grades and ratings, and in a number of the studies, the association has been found to be quite powerful, with correlations ranging up to .90. The fourth section of this document discusses the implications of the findings, concluding that the widely-held belief that grades and ratings are unrelated is a myth, relying for its support on studies conducted more than 20 years ago--studies that are weak in design and execution, and sometimes less than candid in reporting the data. (Author/NHM)

GRADING STYLE

AND

STUDENT EVALUATION OF FACULTY

Robert Powell

William Rainey Harper College

April, 1975

This paper discusses the association between student grades and student ratings of faculty. It is organized as follows: the first section reviews the findings of a 1973-74 study of the ratings of Harper College English teachers; the second section reports the findings of a just-completed 1974-75 follow up study; the third reviews the literature of the field to determine if Harper College results replicate those of other colleges; the fourth discusses the implications of the findings.

The 1973-74 Study

The original study, made at the end of the fall, 1973 term, appeared in March 1974 under the heading "Evaluation and Student Grades". It showed strong associations between student grades and teacher ratings for the 18 full time Harper English faculty members. The coefficient of correlation between grades and ratings as computed by a statistical formula called Spearman's rank order was a high .73. The chances of the finding being accidental was determined to be less than one in 100.*

The strength of the association is illustrated by the following data. The 5 teachers who received the highest mean student ratings (above 4.19 on the 5 point rating scale then used) had given an average of 32% A's and 37% B's to their students. The 5 teachers who ranked lowest (below 3.93 on the 5 point scale) had given an average of 9% A's and 23% B's. The mean grade point average assigned to students by the 5 highest ranking teachers was 2.90, just below B on the 4 point scale. The g.p.a. of the 5 lowest ranking teachers was 2.06, just above C.

The evidence pointed to a powerful relationship between grades and ratings. It suggested that those teachers who insisted on conservative grading standards might be at a very considerable disadvantage in competing for merit raises, promotions and sabbaticals if student evaluation of faculty continued to play a role (as the Harper Board of Trustees insisted it should) in the college's faculty evaluation system. The report thus suggested that computer-summarized faculty rating scales were perhaps of questionable validity,

*For an explanation of the statistical notations used in this paper see "A Statistical Note" -- Appendix E.

in spite of what was assumed to be massive evidence to the contrary, and that their continued use in a competitive faculty evaluation system might set up a jockeying for position among faculty members that could affect the standards of the college and could lower rather than raise the quality of teaching.

The 1974-75 Study

THE SAMPLE

The present study attempts to determine whether last year's findings would be replicated with a new and standardized teacher evaluation form, the nationally distributed University of Illinois C.E.Q. form having replaced the Harper committee-created form used last year. The conditions of rating were also more controlled. Last year, though the student ratings were anonymous, the instructor administered his own ratings and turned them in to the Division office. This year the anonymous ratings were placed in sealed envelopes by students, and the instructor did not see them until the semester was over. In addition, this year a statement was read to the student telling them that the ratings would be used as evidence for promotions and pay increases. In last year's ratings a few students might not have been aware that the results would be used for personnel purposes. The more rigid controls are important. Research studies by Aleamoni and Hexner (Eric ED 081405, 1973) and others havedemonstrated that evaluations tend to be higher when (1) the instructor administers them and (2) when the students are aware that they will affect the instructor personally or professionally.

Of the 18 full-time English teachers in the college 16 made available their confidential computer printouts summarizing the results of student evaluations taken near the end of the fall term in December, 1974. In all, printout of 35 of 40 evaluated sections were voluntarily made available and are included in the study. The college evaluation system requires that teachers be evaluated by at least two of the four or five sections they teach each semester. The division chairman selected the classes to be evaluated. Most of the teachers are thus represented in this study by two classes, but two teachers were evaluated by three classes and one by four. One teacher is represented by only one section.

THE DESIGN: COMPARING HIGH AND LOW GRADERS

The 16 English teachers in the study fall into two separate and easily identifiable groups according to their habitual patterns of grading. Column 1 of Table A shows the mean grade point average given by each of the teachers to all their students in all classes in the fall term of 1973. Column 2 shows the same data for fall 1974. Column 3 is an average of the first two columns. It is the teachers grading style index and is the basis of the 1 thru 16 rankings shown in Column 4. Columns 5 and 6 show the percentage of A's given by each teacher in fall of 1973 and fall of 1974 respectively.

The two groupings are apparent from the table. "High Graders", teachers 1 through 8, assign average grades from 2.50 upward, the top of this group having grading style indexes just above and below B. They are generous with A's, the median for the group being 29%. "Low graders", teachers 9 through 16, assign mean grades of from 2.50 downward to just above and below C. They are stingy with A's, their median being 12%.

The department has low turnover. The teachers are experienced. The grading pattern changed little between 1973 and 1974, even though an "N" (not completed) grade counting as 0 in teacher grade point averages was added to the A through F system in 1974. A check of grades for 1972 and 1971 reveal the same patterns, though the departmental grade point average has risen. High graders remain high and low graders low. The only movement between groups occurs at the very bottom of the high group and the top of the low group.

The grading style index is not influenced strongly by the type of course taught or by time of day the course is given , though literature courses and other electives tend to be graded somewhat higher. The department mean average for evaluated literature courses is 2.70 and for English 101, 2.52. All the teachers usually teach three of four composition courses and one literature. There are no common final exams. Though some sections naturally tend to be of higher ability than others and would tend to receive higher grades, their effect on the grading style indexes is thought to be equal for all. Grades are based exclusively on the instructors own opinion of what the student has earned. Habitually that opinion differs between those who belong to the "high grader" group and those who belong to the "low."

FINDINGS:

Item 9 on the C.E.Q. asks the students to rate the instructor overall performance on a 6 point scale: 1-very poor, 2-poor, 3-fair, 4-good, 5-very good, and 6-excellent. The computer adds all student ratings in each section and prints a mean section rating. These section means become a statistic on the charts used by peer and administrative evaluators.

Table B shows the 35 section means in two columns. The left column lists the 18 courses taught by the 8 high graders. The right hand column shows the 17 courses taught by low graders. The median section rating of the high grader was 5.32, the median score of the low graders was 5.00. Only one class taught by a low grader reached the median of the high graders. The average section of the high graders was 5.22, the average of the low graders 4.78. Every high grader except one placed at least one section at 5.22 and above, and the one who did not make it was close, at 5.17. Four sections taught by low graders reached that level.

The students were asked to place the grade they expected to receive in the course on the rating form. An expected grade mean could thus be computed for each section. Actual final grades assigned by the teacher at the end of the semester were also available. The average mean section expected grade for the low graders was 2.86, for the high graders, 3.13. The average mean section final grade for the low graders was 2.33; for the high it was 2.80. It is significant that the four sections of low-grading teachers that reached the high-grader mean of 5.22 show both expected grades and final grades well above the teacher's usual pattern. The average expected grade for those four sections was 3.07 and the average final grade, 2.65. The final section grades are 40% to 50% of a grade level above the teachers' grading style index.

Coefficients of correlation computed by the Pearson product moment method were:

Mean teacher ratings in 35 sections with:

Student expected grade mean:	.49
Student final grade mean:	.43
Teacher grading style index:	.58*
English 101 sections only with expected grades	.54
Literature & electives with expected grades	.46

The Spearman rank order correlation between teacher grades and ratings in each teacher's highest rated class are shown on Table C. The rank order correlation between mean section ratings and teacher's grading style index is .75.* Between the ratings and section final grades, it is .79.

The 1973-74 study findings were obviously replicated in the 1974-75 study. Unless the Harper correlations are a one in one hundred statistical accident, grades, or whatever grades symbolize, have a very important association with teacher ratings, perhaps accounting for one-third to two-thirds of the differences in teacher ratings.

*Of the five evaluated English sections not made available to this study, four were taught by low graders. Their grading style indexes and rating ranges are known. If it had been possible to include them in the study they would have increased the grading style index correlations.

Conflicts With Expert Testimony

The evidence from the replicated studies thus indicates that in the English area of Harper College there is a continuing relationship between the grades a student receives and the ratings he gives his instructor at the end of the course. The findings, however, run counter to what expert testimony has predicted would be found. In the past several years the college has brought in no fewer than six outside consultants to help it establish a system of student evaluation of faculty. All seems to have ignored or played down the effect of grades on teacher ratings. In doing so they were supported by a large body of literature, their own and others', that repeatedly states that no such association exists.

For example, W. J. McKeachie, one of the consultants brought to Harper, said in 1973 in the Proceedings of the First Invitational Conference on Faculty Effectiveness as Evaluated by Students:

> "The classic research on most aspects of student ratings of instructors was carried out by Henry Remmers and his students at Purdue. His results are still largely unchallenged by more recent research. Among the factors which did not significantly affect ratings were such student characteristics as:
>
> Veteran/non-veteran status
> Age
> Sex
> Class standing
> Grade in Course (However when the top students achieve more than expected they rate the course higher, and when the poorer students do better than expected they rate the course higher.)"

Kenneth E. Eble, another Harper consultant, says in his 1972 book, Professors as Teachers:

> "Scrutiny of thousands of questionnaires at perhaps the easiest point for testing the popularity hypothesis--the correlations between favorable grades and favorable responses--repeatedly shows no correlation."

Professor Eble, the former director of AAUP Project to Improve College, is probably the best known of the contemporary authorities on student evaluation of faculty.

Charlotte Epstein, writing in the April, 1974 issue of the Community and Junior College Journal asks a question:

> How do the perceptions of faculty compare with the findings of scholarly research on the validity of student evaluations?

She took a faculty poll at her community college to find the answer. She reports:

> In most cases faculty attitudes do not agree. Research findings, for example, do not support the faculty view that student ratings are affected by grades, class size, or whether or not the student is majoring in the discipline. Nor do the faculty seem aware of a body of research which cites student ratings as remaining unaffected by the sex of the instructor or the students' grade point average.

The testimony of authority has thus been quite strong in this area. Richard I. Miller in his book, Developing Programs for Faculty Evaluation,(1974), the most complete and scholarly work in its field, comes to the conclusion that grades and ratings are only marginally related, if at all, quoting the previous review findings of Costin, Greenough and Menges of the University of Illinois, published in Review of Educational Research in 1971. Lawrence Aleamoni, still another consultant at Harper, said in an address delivered at the Symposium on Methods of Improving University Teaching held at the Israel Institute of Technology, Haifa, Israel in 1974:

> In almost all the studies cited in Costin, et al. (1971) and by investigators such as Guthrie (1954), Remmers (1960) and Weaver (1960) little or no relationship has been found between a student's grade and faculty rating. In fact, the positive correlations seldom exceed .30. The evidence, therefore, indicates that students do not necessarily rate an instructor or course based upon the grade they have or are about to receive.

Mr. Aleamoni is with the Measurement and Research Division of the Office of Instructional Resources of the University of Illinois, publishers of the Illinois Course Evaluation Questionnaire (CEQ).

It is no wonder that authorities in the field, and administrators who have been in contact with them sometimes express irritation at those faculty members who, operating from a gut feeling, insist that a quantitative evaluation system forces them to play to the wishesof those students who are least ambitious and least able. Professor Eble, who is both authority and administrator, expressed such irritation when in the January, 1974 issue of College English, he attacks an article by Evelyn Kossoff, which had appeared in the winter 1971-72 edition of The American Scholar. Ms. Kossoff, whom he calls a "former English teacher", had criticized ratings from a philosophical rather than an empirical point of view. After calling her article another of the "eruptions of ignorance" that he keeps confronting in respectable places, he says:

> The basis of information from which Ms. Kossoff's article proceeds is (1) "not long ago I saw a questionnaire," (2) "another widely circulated evaluation questionnaire," and (3) two survey articles in 1953 and 1963 general reference works. These sources offer as little information about evaluation as "there flashed through my mind the picture of one professor who..." (Ms. Kossoff's words) affords about the nature of effective teaching. It is as if she set out to question the validity of current cancer research by citing a pamphlet picked up in a chiropracter's office and an article in a 1953 encyclopedia... All this is bad enough as measured by any standards of scholarship, but it is worse when one considers that a writer working within a University community might come across one or more of the following:

Professor Eble then goes on to explain that there were 50 items that might have come to Ms. Kossoff's attention, including "the existence of the University of Washington office of student evaluation since 1925; and the fact that I (Eble) had been on her campus (U. of Kentucky) the previous October talking with a campus-wide audience about evaluation." He continues: "That is why I turn to willful ignorance as an explanation of this kind of imperviousness to information on a subject important to college teaching...The examined life is held up as a scholarly ideal as long as it stops short of examining teaching."

One can well understand Professor Eble's frustrations, as he expressed them in the College English article, which he entitled "What Are We Afraid Of?" He has worked long and hard in the field, making some progress, as testified to by his statement in Professors as Teachers: "My recent inquiries suggest that the use of systematic student ratings has greatly increased since 1966."

Collecting the Data

The fact remained that in the Harper College English courses, grades and ratings were so closely connected that the wisdom of continuing to use quantified evaluations could be called into question. Even though expert testimony pointed to error or statistical accident as the cause of the Harper findings, it was thought best to look at the past research done in other colleges for clues as to why the Harper findings were different. It was decided to do a thorough job, to avoid the easy habit of picking up studies in chiropracter's offices or using examples of the "there flashed through my mind a picture" type. It was decided to do the most thorough job yet attempted in this field--to locate and summarize every original source research study published since 1930 that focused in whole or part on the student grades-teacher ratings topic. It was decided to look closely at the research, not just at the conclusions the researcher reached, but to study the intention, sample, design and execution of the work as well.

The collections of two large university libraries were searched. An ERIC computer search was run. More than 200 review studies were read for their references and their bibliographies. More than 75 studies were xeroxed and summarized. Many turned out to be secondary sources or student achievement studies, but 41 seemed to meet minimum requirements, that is the author had looked at a body of student ratings in a planned way for the specific purpose of determining if grades and ratings were related, and had said something about the size of his sample, his method of examining the data or the results of his investigation.

It became plain early in the search that many studies listed in the review literature as showing that grades and ratings were unrelated had barely enough substance to make the list of 41. Their authors had made no serious attempt to compare one teacher's results with another or they had used as samples teaching assistants who had no say in what grade the student earned or they had used

questionable research designs. These weaknesses were not apparent when one read only the researcher's conclusion, but became obvious when one went deeper into the study.

Table D lists all of the 41 empirical studies found in the search. It is non-selective. No study has been omitted. Those studies that seem to meet minimum research standards are marked with an asterisk. The studies are arranged chronologically from 1930 onward in Column 1 of the table. Column 2 shows the number of students, teachers or sections involved in the study. Columns 3, 4 and 5 show the strength of the grade-rating associations found by the researchers. If the authors of the study concluded that the correlation between grades and ratings was negative, nil or negligibly positive, the results are entered under Column 3. If weak to moderate associations were found, the results are found under Column 4. Finally if marked or strong associations were found, or if the author believed the assocation he found to be quite important, it is entered under Column 5.

It's quite obvious from a glance at the table that those who maintain that research has shown that there is no relationship between student grades and student ratings of faculty seem to be right up to a point -1953 to be exact. But this paper will hereafter show that they were not right even to 1953. It's quite obvious that they are not correct from 1953 onward. Of the 28 studies conducted since 1953, six showed negligible correlations, 10 showed low to moderate correlations and 12 showed marked to strong correlations.

Because of the apparent conflict between what Professor Eble and others who work in this field have said about the relationship of grades to ratings and Table D, it will be worthwhile, though time-consuming to look more deeply at every one of the studies so that better judgments can be made about what they signify. All 41 are summarized on the following pages. The summaries begin with the studies, all made from 1953 onward, that are listed in Columns 4 and 5 of Table D and show at least a low relationship between grades and ratings. After that the studies that showed negative, nil or negligible correlations are summarized. All but six of these were conducted before 1953.

It should be remembered that some of the studies were concerned not only with grades and ratings, but with other facets of student evaluation of faculty as well.

In most instances, the findings on effect of class size, time of day of class and the like have been omitted. Focus is on the key issue--the effect of grades on the validity of the scales.

Some studies were surely missed in the search, and several listed in bibliographies could not be found. Those unpublished studies that must lie in filing cabinets at various colleges could not of course be included. Still, it is felt that the 41 studies are the most complete collection yet assembled and give a comprehensive picture of empirical research up to the appearance of mid 1974, periodical indexes.

Studies Showing Positive Grade-Rating Correlations

The first research to show a relationship between student grades and teacher rating was described in the article by A.M. Anikeef, appearing in 1953. It followed a quarter century of studies that unanimously insisted that grades were not important to ratings. The very considerable prestige of psychologists H.H. Remmers at Purdue and E.R. Guthrie at the University of Washington supported the no grade-bias position. Anikeefs' was the watershed. Since it appeared the great bulk of the original research has shown that grade bias exists. The research that has not will be shown in the next section to be of questionable believability. Anikeef's study and the Column 4 and 5 studies that followed it are summarized below. The numbers that precede the author's names and the study title identify the study's position on Table D. Anikeef's is the third study from the year 1953.

1953-3: A.M. ANIKEEF: "Factors Affecting Student Evaluation of College Faculty Members." Journal of Applied Psychology, 37, No. 6, 1953.

Anikeef studied 1500 ratings of 19 instructors in the School of Business and Industry at Mississippi State College. He found a correlation of .73, significant at the .01 level, between grading leniency scores (the mean grade point average assigned by the instructors) and the ratings of their instructors by freshmen and sophomores. For junior and seniors he found a correlation of .43, which he did not claim to be statistically significant -- a correlation of .48 being needed if significance is to be claimed when working with the number of teachers in his study. The combined freshmen-senior correlation was .53. Anikeef concludes that 53% of the variance in freshmen-sophomore ratings and 25% of the combined freshmen-senior ratings could be attributed to grading leniency.

Comment: The design Anikeef used is similar to the one used in the Harper English studies. The Spearman rank order correlations found are similar. A merit pay system was in use at Mississippi State too.

1960-1: CARL H. WEAVER: "Instructor Ratings by College Students" Journal of Educational Psychology, 51, No. 1, 1960.

The article reports a study of 699 student ratings in 39 sections of history, English, personnel and speech taught by 12 different instructors at Central Michigan Unvirsity. The teachers were not compared as in the Anikeef study. Instead, expected grades listed by students on the Purdue rating forms were pooled. Expectant A's gave mean ratings of 96.10, B's 94.56, C's 91.15, D's 84.63. The differences were significant at the .001 level of confidence. The author suggests that grade bias is of real importance in interpreting ratings.

1964: PAUL P. ECHANDIA: "A Methodological Study of Factor Analytic Validation of Forced Choice Performance of College Accounting Instructors." Dissertation Abstracts, 1964 (25) (4) 2605-2606.

Studying 546 accounting students of 16 teachers at New York University, Echandia found that students who recieved higher grades in the course rated their teachers significantly higher on factors concerned with course organization and lucid exposition. Motivational factors were not significantly correlated with grades. No correlation figures are given in the abstract.

1965-1: R.E. SPENCER AND W. DICK: Reported in "The Illinios Course Evaluation Questionaire: A Description of Its Development ans a Report of Some of Its Results." by Lawrence Aleamoni and R.E. Spencer in Educational and Psychological Measurement, 33, 1973.

Sample: 600 students in two courses at Pennsylvania State rating their instructor using the Illinois Course Evalution Questionaire (CEQ) developed by Spencer. Whether the two courses had more than two sections or two instructors is not stated. Finding: "Course grades and rating scores did correlate significantly (even though magnitude of the correlation was small) with all the subscores except the instructor rating." Exact correlations are not given.

1965-2 R.E. SPENCER AND W. DICK (1965-2). Same sources as 1965-1 above. Sample: 160 students in 12 sections of Speech 101 at Pennsylvania State, using the CEQ form. Findings: grades on six speeches and three tests all correlated impressively with ratings -- .85 for speeches, .86 to .91 for each of the tests.

Comment: These two studies were apparently reported first in the 1965 edition of the Manual of Interpretation for the CEQ by Spencer and Dick. The 1972 edition of the manual, by Lawrence M. Aleamoni, seems in its 64 pages to contain no specific reference to the Spencer and Dick studies or to make any mention of a relationship between grades and ratings. Aleamoni and Dick do, however, in their 1973 article in Educational and Psychological Measurement say, "It can be seen, then, that in some courses, student opinion about the course is highly related to success in the course." The CEQ form is used by the students at Harper College.

1966-2: CLIFFORD T. STEWART AND LESLIE F. MALPASS. "Estimates of Achievement and Ratings of Instructors." The Journal of Educational Research, Vol. 59, No. 8, 1966.

Sample: 1975 students rating 67 instructors teaching 53 courses at the University of South Florida. Findings: "Highly significant relationships were observed between estimated course grades and ratings of instructor-variables." These included strong associations between expected grades and approval of the teachers grading policy. The relationships were significant well beyond the .001 level.

1969-1: B. DAYLE WALKER. An Investigation of Selected Variables Relative to the Manner in Which Population of Junior College Students Evaluate Their Teachers. Dissertation Abstracts, 1969, 29 (9-B), 3474.

According to the abstract, 1447 students of 30 teachers at Lee Junior College rated their teachers on the Purdue Rating Scale. No statistical correlations are given, but the abstract says, "Students tend to rate teachers in the direction of their stated anticipated grades."

1970-1: J. RUBENSTEIN AND H. MITCHELL:"Feeling Free, Student Involvment and Appreciation." Proceedings of the 78th Annual Convention of the American Psychological Association, 5, 1970.

Sample: 1655elementary psychology students at Purdue in 60 sections.

Results: Class grades earned up to the date of the rating correlated .14 to appreciation of instructor and .30 to appreciation of the course. Final course grades correlated .09 to appreciation of instructor and .44 to appreciation of course.

1971-1: DAVID S. HOLMES: "The Relationship Between Expected Grades and Students' Evaluations of Their Instructors." Educational and Psychological Measurement, 31, 1971.

Holmes studied ratings by 1539 students in 7 large lecture classes with enrollment of more than 100 at the University of Texas. Grading was by objective exams. He found statistically significant but small relationships between expected grades and two of the three rating subscales. The series of items gathered under the heading "Student Stimulation" were all associated with expectant grades, as were most of the items uder the heading "Interaction-Evaluation." However, only one item under the heading "Instructor Presentation" was found to be moderately related to grades. The mean amount of variance shared by grades and key evaluation items was found to be 5% and the maximum 13%.

1971-2: RICHARD G. WIEGEL, E.B. OETTING AND DONALD L. TASTO: "Differences in Course Grades and Student Ratings of Teacher Performance." School and Society, 99, 1971.

At the beginning of their study Wiegel and his associates say ". . . reports dating as far back as 1928 have shown there to be only a negligible relationship between course grades and the teacher performance evaluations." They then describe a study of the ratings of 4 teachers by 331 students in 7 psychology sections at Colorado State College. They found a strong positve correlation, significant at the .01 level, in some classes and not in

others. Pooled ratings for the 7 sections showed positive correlations, alsosignificant at the .01 level. They conclude ". . . even though large correlational studies indicate that students' grades and evaluation of the teachers are not importantly related, this relationship should not be dismissed lightly. The effect is likely to be idiosyncratic for both teacher and course, and should be considered in planning of interpreting teacher evaluations."

1972-1: R.B. BAUSELL AND JON MAGOON: "Expected Grade in a Course, Grade Point Average and Student Ratings of the Course and the Instructor." Educational and Psychological Measurement, 32, 1972.

Bausell and Magoon examined 12,000 ratings taken university-wide at the University of Delaware in fall, 1969. They report ". . . the present study found strong consistent biases in both instructors and course ratings which can be traced to (a) the grade the student expects to receive and (b) the discrepancy between the students' expected grade and his G.P.A.. The relationship between the G.P.A. and rating alone is negligible, and should not be considered an important source of bias." The coefficient of correlation between expected grades and ratings was found to be .62 and between discrepent grade and ratings .53, significant beyond the .001 level.

1972-3: ALAN NICHOLS AND JOHN SOPER: "Economic Man in the Classroom." Journal of Political Economy, 80, Sept-Oct, 1972.

Nichols and Soper studying 339 social science sections at Central Michigan University in fall, 1970, compared section mean expected grades and section mean instructor ratings and found a correlation coeffcient of .53. They suggested that the university's grades were again on the rise following the introduction of a compulsory evaluation-by-student system. They also suggest that by raising the mean grade point average of a section a half grade level, an instructor could expect a half grade level rise in his mean section ratings.

1972-4: W. ROBERT KENNEDY: "The Relationship of Selected Student Characteristics to Components of Teacher/Course Evaluation Among Freshmen English Students at Kent State University." Paper & Symposium Abstracts of the 1972 Meeting of the American Educational Research Association,

Sample: 549 freshmen in English 160 at Kent State University, Fall, 1970.

Findings: grade point averages, final grades and expected final grades all correlated significantly with teacher ratings. Student ability, as measured by A.C.T. scores, did not seem to be related to teacher rating.

Comment: This is the third study available only in abstract and specific figures arelacking. The tone of the abstract suggest quite strong associations between grades and ratings, but it and the other two have been put under the "slight to moderate" heading on Table D because of uncertainty. In none of the three abstracts does the summary suggest that the relationship is negligible.

1972-5: DAVID S. HOLMES: "Effect of Grades and Disconfirmed Grade Expectancies on Student's Evaluations of Their Instructor ." Journal of Educational Psychology, 63, No. 2, 1972.

In an introductory psychology class of 97 students at the University of Texas course grades were based on four objective tests. After completing three of the tests each student knew exactly what grade he had earned up to that time. Student wrote the grade they expected to receive in the course on the final test paper, after being promised that the expected grade would in no way influence the final grade. When the students returned to collect their final exam and learn the final grade, half of those who had both expected and earned A's and half of those who had both expected and earned B's were told that their final grade was one level lower than they expected. The other half was told the truth. They then completed teacher rating forms before they all got their finals back and learned there had been an experiment. No difference was found between the ratings given by A and B students, but those whose grade expectancy had been disconfirmed rated the teacher

significantly lower on teacher preparation, lecture, coherence, use of examples, ability to evaluate, value to the students and test clarity.

1973-1: BARAK ROSENSHINE; ALAN COHEN AND NORMA FURST: "Correlates of Student Preference Ratings." Journal of College Student Personnel, 14, May, 1973.

The study was of 1200 daytime classes in all the schools and colleges of Temple University in spring, 1970. The methods of administration and the reasons for administering the scales are not discussed. The authors found correlations of .09 to .27 between expected grades and two questions that asked the student to compare the instructor and the course with others. A four point rating scale was used. They found no correlation between grade point averages and ratings. They conclude that the effect of expected grades on ratings, though statistically significant, is low.

Comment: The Rosenshine rating form asked the students to rate teachers on 23 items measuring classroom style and behavior. Of special interest to English teachers are the three items that showed the lowest correlations with student appreciation of the class and of the instructor. They are:

Criticism of papers was helpful to the students	.26
Instructor used assigned papers as an aid to learning	.21
Instructor criticized student responses in destructive way	-.16

Low correlations were also found for the following variables: a) independent projects and papers, b) class participations, c) creative thinking, d) application and appreciation were important for the final grade.

Much higher correlations were found for the following items:

Instructor's main emphasis was on student's learning	.40
Grading in the course was fair	.46
Instructor's main emphasis was on having the students enjoy the course	.50
Instructor was enthusiastic	.57
Instructor's presentation was clear and understandable	.62 (the highest of 23)

The authors comment that though criticism of papers is often cited as being important for college teaching, its importance was not borne out by the data collected in the study.

These findings seem to suggest that the following teacher behavior may not be conducive to high teacher ratings:

1) assigning complex, symbolic hard-to-explain readings
2) emphasizing learning over student enjoyment
3) being unfair in grading -- grading harder than one's peers
4) counting papers toward the final grade, particularly if the papers require creative thinking and application of knowledge
5) writing negative criticism on papers

1973-2: RICHARD R. PERRY AND REEMT R. BAUMANN: "Criteria for the Evaluation of College Teaching: Their Reliability and Validity at the University of Toledo." Proceedings, The First Invitational Conference on Faculty Effectiveness as Evaluated by Students, ed. Alan L. Sockloff, Temple University Measurement and Research Center, 1973.

Perry and Baumann - analyzing 900 students ratings at the University of Toledo in Spring 1972 - found correlations of up to .78 between class mean expected grades and class mean ratings, with an average of .42 for all levels of the institution. They said of the rating scales "the indictment of the validity is very strong; what the correlations reveal is that variations in course ratings is accounted for to the extent of 30 to 60% by the grades assigned. . . this problem must be resolved in some fashion before one can build a reasonable case for validity."

1973-3: JOHN A. CENTRA AND ROBERT L. LINN: "Student Point of View in Ratings of College Instruction." An Educational Testing Service Research Bulletin, October 1973. ERIC Document 089581.

The study was of 300 randomly selected students from 402 classes in 5 colleges. Grades were found to be "moderately" related to ratings though not in all classes. No specifics were given. The authors say that their findings underscore the importance of context of the course in determining ratings.

Comment: This 1973 Centra and Linn study does not seem to be mentioned in the 1974 ETS SIR (Student Instructional Report) manual of interpretation

or in the portfolio of materials ETS distributes to advertise the SIR rating scales.

1973-4: ROLF MIRUS: "Some Implications of Student Evaluation of Teachers." Journal of Economic Education, 5, No. 1, 1973.

Mirus studied 122 course evaluations (unstated number of students) at the Faculty of Business Administration and Commerce at the University of Alberta in 1971-72. He compared mean section expected grades with mean section instructor ratings, section by section. Finding a correlation of .85, he states, "There is a strong indication that the expected grade is a major determinant of the professor's grade. . . A professor who, compared to his colleagues, makes the class expect a 1.00 point higher grade can improve his own evaluation .85 of a point." Mirus suggests that the higher coefficient or correlation between grades and ratings found in this study as compared to the Nichols and Soper is because the career orientation of the Alberta students makes them more responsive to grades. Mirus asserts that an updrift of institutional grades can be expected as a result of the evaluation system. A statistically significant higher average grade was reported in 1972 as compared to 1971.

1973-5: K.L. GRANZIN AND J.J. PAINTER: "A New Explanation for Students Course Evaluation Tendencies." American Educational Research Journal, 10, No. 2, 1973.

The authors gave first day of class "expectation" questionaires and the last day of class "rating" questionaires to 637 students in 17 courses offered in 11 different departments at the University of Utah. Among correlations found to be significant at the .001 level of confidence are:

Course rating to expected grade .21

Course rating to final grade .15

Course rating to expected grade change - higher rating at the

end than expected at beginning .18

Instructor rating and expected grade .16

Instructor rating to expected grade change .14

No significant correlation was found between student grade point averages and ratings. Final grades (as contrasted to expected grades) correlated .09 with instructor rating.

1973-6: ALLEN J. SCHUH AND MICHAEL A. CRIVELLI: "Animadversion Error in Student Evaluations of Faculty Teaching Effectiveness." Journal of Applied Psychology, 58, No. 2, 1973.

A class of 85 students in a required business administration degree course in industrial relationships were asked to rate their instructors immediately after he has returned their midterm exams. The instructor left the room while the ratings were administered. Ratings were found to be associated with midterm grades beyond the .001 level of significance.

1974-1: C.D. CORNWELL: "Statistical Treatment of Data from Student Teaching Evaluation Questionaires." Journal of Chemical Education, 51, No. 3, 1974

Sample: An unstated number of students in 101 different chemistry lecture sections taught by 70 different lecturers in 20 different institutions. The data was collected by a committee on Undergraduate Teaching of the American Chemical Society. Findings: Statistically significant but weak relationships were found between grades and ratings. The research estimates that the grades accounted for 1% of the variance in ratings.

1974-2: WILLIAM M. BASSIN: "A Note on the Biases in Students' Evaluations of Instructors." The Journal of Experimental Education, 43, No. 1, 1974.

Mean grade point averages given by 64 teachers at the University of Maryland were compared with the mean ratings given them by students. Bassin found an overall coefficient of correlation between grades and ratings

of only .10. However he found that this minor correlation was associated with a major effect on teacher rankings. The average teacher teaching a quantitative course, giving a grade point average of 2.0, ranked at the 30th percentile in student rating of lecture quality. The average teacher teaching a quantitive course, but giving a 2.5 grade point average, ranked at the 62nd percentile in student rating of lecture quality.

Examination of 22 of the 28 studies made since 1953 shows clearly that grade-rating correlations do exist and that the associations between grades and teacher ratings can be quite powerful. Before reaching a conclusion about the relevance of these studies to the Harper English study it would pay to look rather carefully at and comment on the 19 studies published since 1930 that have led many people to believe that grades and ratings are unrelated.

Studies Showing Negligible Grade-Rating Correlations

Column 3 of Table D shows that the authors of 19 of the 41 studies have concluded that the ratings of teachers are not biased by the grades the teacher gives. The 13 studies made before the Anikeef study of 1953 were unanimous in taking this position. Since 1953, six of the 28 studies have supported the no-bias position. Detailed summaries and comments follow.

1930: H. H. REMMERS: "Two What Extent Do Grades Influence Student Ratings of Instructors?". Journal of Educational Research, 21, 1930.

Remmers of Purdue popularized the use of the rating scales in colleges. This study, first published in a shorter form in 1928, made use of 17 classes. Seven were high school classes taught by practice teachers. Ten were college classes taught by four different instructors. Data was collected as follows. When the students completed the rating forms the teacher read off the names of those students ranking in the top half of the class, asking them to put an X on the form. Remmers then correlated non-X and X ratings within each class. Some classes showed positive grade-rating correlations; others showed negative correlations. When he averaged all the correlations from 17 classes, he found a mean correlation of only .070 "at the most". He therefore concluded "...for the average instructor and the average student there is practically no relationship between a student's grade and his judgement of the instructor as recorded in the Purdue Scale for Instruction."

Comment: The study is of course a collection of 10 separate "within class" studies of the classes of four college teachers, subjects, and methods unstated. Remmers did not compare teachers, even though rating scales by their very nature do compare teachers. It is not to be expected that single within-class studies of this type will always tell something about the relationship of grades to ratings. They can not detect differences in grading style, nor will they in all cases

show positive correlations between abilities and ratings, even when powerful associations between the two exist throughout a department or a college. The low ability and the high ability students of a "high" grading teacher may be equally happy with him and give him equally high ratings, since all are exposed to the same grading style and equal numbers from both groups may be earning higher grades than they are accustomed to making elsewhere. Likewise low and high ability groupings from classes of "low" graders may contain roughly equal numbers of students who are experiencing more trouble with grades than they are accustomed to, or that their peers are experiencing in comparable classes. The result could be equally low ratings from both groups and again a lack of positive correlation between grades and rating. If Remmers had pooled results across sections, then drawn coefficients of correlation for the entire group, he would perhaps have detected the small correlation that has sometimes been found in studies in which student abilities, as indicated by their grade point averages, have been correlated with teacher ratings. Among studies that have not found the gpa correlation are study 1972-1 (Bausell and Magoon) and study 1954 (Clark and Keller). Among those that have found small positive ability correlations is study 1950-2 (Elliott). See 1971-[illegible] (Wiegel) for a small study of 4 teachers, similar to the Remmers study but with quite different results, when A, B, C, D grades were pooled across sections.

1934: J. A. STARRAK: "Student Rating of Instruction." Journal of Higher Education, No. 5, 1934

Starrak reports that 40,000 scales have been taken at Iowa State College since 1928. He gives no details of the method of collection or size of the sample used to reach the conclusion that the correlation between grades and ratings is only .15. This correlation he believes is small enough to be disregarded.

1936-1: J. D. HEILMAN AND W. D. ARMENTROUT: "The Rating of College Teachers on Ten Traits by Their Students." The Journal of Educational Psychology, 27, 1936.

The authors studied ratings taken in 50 classes taught by 46 different

teachers at the Colorado State College of Education in Spring 1935. Average class size was 42. Teachers apparently administered their own ratings and voluntarily turned them in. The authors found a severity of grading score for each teacher by averaging all grades he assigned in the 1935-36 school year. The severity of grading scores were then compared with mean section student ratings. The correlation found was -.042. The authors therefore conclude that there was no relation between student grades and teacher ratings. They comment at some length, however, on the difficulties individual instructors had in interpreting the meaning of the scales. Average section standard deviations were very high. One instructor, for example, was found to have a standard deviation of 27.30 on the 100 point scale, suggesting such a wide scatter of student opinion as to deny the existence of a center.

Comment: The Heilman and Armentrout study is well designed to detect the influence of grades on ratings. It is an admirably detailed study. Though the teachers took their own ratings, somewhat weakening its believability, it seems to this reviewer to be the only study in the literature which truly supports the conclusion that the grades a teacher gives and the ratings he receives can be unrelated. One may, however, wonder about the size of the standard deviations found. Their size may indicate that something was wrong with the scale or with its administration.

1936-2: MILTON L. BLUM: "An Investigation of the Relation Existing Between Students' Grades and Their Rating of Their Instructor's Ability To Teach." Journal of Educational Psychology. 27, 1936.

This is a study of 57 students in two 8-week summer psychology classes taught by the same teacher at City College of New York. Blum found no relationship between expected or final grades and instructor rating. Forty of the 57 students were expecting A's and B's at the time of the

rating. Sixteen were expecting C's, one a D.

Comment: Lacking an experimental design such as those found in the single teacher "within" class studies of Schuh and Crivelli (1973-6) and Holmes (1972-5), this study merely adds two more within class studies to the 10 found in Remmers (1930). They show that a teacher need not always expect to find positive grade rating correlations within his own classes. The study is of no value in telling the teacher how the grades other teachers give affect his ratings.

1949: H. H. REMMERS, F. D. MARTIN, AND D. N. ELLIOTT: "Are Student Ratings of Instructors Related to Their Grades?" Purdue Studies in Higher Education, 66, 1949.

The study evaluated 37 graduate assistants teaching the lab and recitation sections of the freshman chemistry course at Purdue. The senior professor who gave the lecture - demonstration was not evaluated. The graduate assistants had little to say about course grades, exams being standardized and departmentally graded. The researchers divided the students into two groups: the plus group consisted of those whose final grades were higher than pre-course placement tests predicted; the minus group received grades lower than predicted. The plus groups were found to rate their instructors significantly higher (.13 to .35) than the minus group. Since the researchers found no relationship between placement test scores and ratings, and since the assistants did not control the grades, they conclude that the connection found is not higher grades cause better ratings but that better teaching causes higher ratings. The authors also gave their attention to the phenomena that Remmers had first noticed 20 years before, the fact that some of the classes in his within-class study of 1928-30 showed a negative correlation between grades and ratings, other showing positive correlations. They

now suggested that some teachers are good at teaching high ability students and poor at teaching low ability students. These teachers, they reason, will receive poorer ratings from the weaker students and thus show positive correlations between grades and ratings. On the other hand, teachers who are best at teaching low ability students will alienate some high ability students and show negative correlations between grades and ratings.

Comment: The Remmers explanation of his findings of 20 years earlier may suggest the interesting possibility that the best method of achieving consistently high mean student ratings would be to teach to the abler students and to see to it that the less able were not disaffected - that is make the weaker students feel successful. See Holmes (1972-5) for a possible explanation of the action of the minus group. Also see Bausell and Magoon (1972-1).

1950-1: DONALD N. ELLIOTT: "Characteristics and Relationships of Various Criteria of College and University Teaching." Purdue University Studies in Higher Education, 70, 1950.

Donald Elliott was Remmer's assistant in the Division of Educational Reference at Purdue. His first study seems to be a continuation in greater detail of Remmers, Martin and Elliott (1949). Freshman chemistry assistants were again involved. Only 9% of the assistants had previous teaching experience. Most did not plan to become teachers. The senior lecturer was not evaluated. The assistants had little to say about the grades. Tests were departmentally designed and evaluated. Elliott found correlations of grades to ratings of only .032 for lab sections and .049 for recitation sections. He did however find a correlation of .24 between ratings and achievement, that is he found that students who got better grades than their pre-course tests indicated they would tended to rate their teachers higher than students who did not achieve as much. Elliott also found a negative correlation between student achievement and teacher knowledge of chemistry. The students who achieved most (as measured by grades higher than predicted by placement tests) tended to be most often in classes taught by teachers who scored lowest on a test of knowledge of chemistry.

Comment: This final finding of Elliott is fascinating. He who knows least teaches best, and gets the highest student ratings. Its significance to the Harper English Department study is unknown, but one may speculate. The problem Elliott faced in this study is the one faced by Remmers, Martin, Elliott (1949) and by all the many researchers who have tried to prove that

student ratings of faculty are related to what the student learns. The problem is that one can never be certain that the student who scored higher on an exam or received a final grade higher than his GPA or aptitude tests indicated he should is rating his teacher higher because he has learned more than he expected or because his grade is higher than he expected it to be. Again, refer to the Bausell and Magoon study (1972-1) for an explanation of the discrepant grade effect.

1950-2: DONALD N. ELLIOTT: (The Second Study Found in "Characteristics and Relationships of Various Criteria of College and University Teaching," Above.)

This study was the second of two undertaken by Elliott as material for his doctoral dissertation. According to Elliott 26,014 ratings of 460 instructors had been collected from 14 Indiana colleges and universities as part of the Indiana College Evaluation Program. He mentions the numbers exactly. However, he says only those ratings taken at Purdue contained information about grades. At Purdue, the instructors, following Remmers plan of 1928-30, asked their upper-half students to put an X on the forms. A total of 3786 ratings (1906 upper and 1880 lower) were then available for comparison. The ratings were grouped, not treated as within - class ratings as in 1928-30.

In all categories except that of the graduate student the upper half students rated the instructor higher than the lower half of class. Sample mean scores from the scale that were found significant at the .01 level of confidence were:

	Upper	Lower
Fairness in grading	89.15	82.40
Presentation of subject matter	75.80	73.35
Simulating intellectual curiosity	75.05	72.65

The lower-half students gave slightly lower ratings for every other item on the 10 point scale. Elliott concludes "....the factor of scholastic success has such a slight effect, albeit the effect is statistically significant, as to be virtually ignorable, particularly when it is recalled that most classes are made up of students of widely varying scholastic success."

Comment: The Manual of Instruction for the Purdue Rating Scale for Instruction by H. H. Remmers and J. A. Weisbrodt (Revised edition, 1965) copyright by the Purdue Research Foundation, contains the following paragraph as its total contribution to the grade-rating controversy:

> "Several questions have been raised regarding other factors that might affect the student ratings of instructors. Remmers and Elliott (16) have answered many of these questions. In a study of the ratings of 460 instructors by 26,014 raters in 10 different institutions of higher learning they found that freshmen rated their instructors no higher and no lower than did seniors, male students rated their instructors no differently from female students, veteran students rated their instructors similarly to non-veteran students, and students in the upper half of the class rated their instructors like those in the lower half. None of these factors had any effect on the ratings by the students."

Someone is mistaken, either Elliott or the editors of the Manual. It is unlikely that two separate studies would start with exactly 26,014 ratings.

1951: EARL HUDELSON. "The Validity of Student Rating of Instructors" School and Society, 73, 1951.

This is a one teacher study with a difference. Hudelson asked his 192 students to rank their former teachers anonymously. He then asked them to give the grades they had received from the teachers. Finding a correlation of only .19 between grades and ratings, he concludes, "Obviously these students could not fairly be charged with letting marks influence their opinions of their instructors as teachers."

Comment: The system of collecting data, somewhat similar to that later used by Voeks & French (1952-1, 1952-2), could lower the positive correlation since it removes from the sample those who were forced to leave school because of the low grades they received. Hudelson is the only researcher found in the literature who provides a scatter-diagram to illustrate the association between grades and ratings. Though he did not give mean teacher ratings for each grade level, it is instructive to the reader to examine the scatter-diagram closely and to do his own arithmetic. If he does so, he will discover that the weak .19 correlation was produced by the following data:

The 38 A students gave mean ratings of 6.8.
The 87 B students gave mean ratings of 6.5.
The 57 C students gave mean ratings of 5.6.
The 8 D students gave mean ratings of 5.6.

All ratings were on a 10 point scale. The teacher who gave an average grade of C to his classes might thus expect to produce average student ratings about 9/10ths of a decile below those who gave an average grade of B. In the 1974-75 Harper English study the average mean final grades of the low graders was 2.33, of the high 2.80, a difference of a half grade level. The average mean ratings of the high graders was 5.22 on the 6 point scale and of the low 4.78, a difference of approximately 8/10th of a decile. The Harper grade-ratings relationship are therefore seen to be approximately twice as strong as those found by Hudelson in 1951 -- results not inconsistent considering Hudelson's method of collecting data and the merit system at Harper. It is obvious, in spite of Hudelson's conclusion, that the 1951 study does not show a negligible association between grades and ratings, but rather shows the opposite. It is possible that Starrak (1934) with his .15 correlation "small enough to be disregarded" also belongs in another column in Table D.

1952-1, 1952-2, 1952-3: VIRGINIA W. VOEKS and GRACE M. FRENCH. "Are Student-Ratings of Teachers Affected by Grades." Journal of Higher Education, 31, 1960.

These three studies, which were specifically focused on the grade-rating relationship, were part of a series on a number of aspects of teacher evaluation undertaken under the direction of E. R. Guthrie at the University of Washington. The research was done in 1952, but publication was delayed until 1960.

Data for the first two studies was collected at spring registration. Students of advanced sophomore or higher rating were asked to nominate teachers who fitted the five categories of the Washington teacher rating scale: very superior, superior, competent, only fair, of less value to me than the others.

The researchers then computed mean ratings for those teachers nominated 20 or more times. They also collected the grades these teachers had assigned

during the preceding two terms.

Study 1952-1: In the first study the researcher drew rank order correlations in three departments between student ratings and 1) percentage of A's and B's and 2) percentage of D's and F's the teacher gave in the preceding two terms. They report:". . . . all the correlations between grades and student ratings were negligible (see Table I). No correlations was reliably greater than zero, even at the 5 per cent confidence level." The essential part of Table I are reproduced below.

Department	No. of Faculty	Correlation of the Ratings Assigned by Students and the Percentage of Each Grade Given		
		A & B	C	D & F
A (Physical Science)	10	.00	-.31	+.04
B (Physical Science)	11	+.60	-.17	-.05
C (Humanities)	13	+.05	-.21	+.36

Comment: It is difficult to understand why the author chose to display the data in the above way.. The mean grade point average of each teacher should have been available. Correlations drawn between mean ratings and mean teacher grades would have been much more useful to the reader. The strong correlation between A's and B's and the ratings in Department B suggests that a rank order correlation based on mean grades could approach the correlation levels found in the Harper English studies. In the other departments, the negative correlations under the C's suggest that a rank order coefficient based on mean grades might produce correlations in the range of .15 to .30. The correlation of .36 under the D & F column in Department C is of little significance on a study that eliminated many D students and lower level C's from the sample by taking ratings only from those who survived to at least advanced sophomore status. The statement "No correlation was reliably greater than zero, even at the 5 per cent confidence level" has little meaning when a study is restricted to 10 to 13 teachers. Coefficients of correlation have to be in the range of .55 to .65 before significance can be claimed with such limited numbers.

Study 1952-2: In the second study the authors compared the highest and lowest rated teachers in each of 10 large departments. They provide Table II to show results. The essential parts of the table follow:

Department	No. of Students	Mean Grade Highest Rated	No. of Students	Mean Grade Lowest Rated
Architecture	81	3.185	151	2.338
Art	126	3.206	90	2.400
Chemistry	79	2.633	239	2.155
Economics	365	2.123	68	2.838
Education	134	2.888	241	2.822
English	165	3.062	62	2.145
Math	100	2.350	45	2.155
Political Sci.	267	2.588	62	2.564
Psychology	97	2.588	439	2.414
Sociology	237	2.477	81	2.222

The researchers comment:

> As Table II shows, the teacher with the highest student-rating in his department usually had given a slightly higher average grade than the teacher with the lowest rating...These differences are very slight and often not statistically significant... analysis shows no reliable difference between the mean grades given by the ten teachers with high student-ratings and the mean grades given by the ten teachers with low student-ratings...in the relatively rare instances in which a teacher with high ratings also gave appreciably more high grades, it is evident that he did not receive higher grades because he gave more than the average number of low grades.

Comment: The comparisons in the above table are perhaps unfair, large-section lecturers, who may not be personally involved in grading, being compared with seminar teachers. However, the table does show that in 9 out 10 departments the highest rating went to the man with the higher grades. In four departments the difference is quite large. The English Department difference, 90% of a grade level, is close to the difference between the highest and lowest rated teacher in the Harper English study. It is difficult to see how the authors could make their generalizations on the basis of the data they display, particularly since the method of collecting ratings would serve to eliminate disaffected low-graded students.

Study 1952-3: In the third study the researchers found 16 teachers who had given the Washington course evaluation questionaire to different sections of the same course, the ratings being taken at least one quarter apart, and who had scored at least three deciles higher on the second administration than the first. They then compared the grades the teacher had given

the first time with the grades they had given when they scored the higher ratings to see if the grades had gone up with the ratings. They supply a complex table that shows whether the difference in grades could be explained by chance. They found that one teacher had given appreciably lower grades to the class that gave him the higher rating. On page 333 they report that two teachers had given higher grades approaching statistical significance (.07 and .01) to the second class. On page 334 they report that only one teacher had given appreciably higher grades the second time. They conclude, "Usually the grades given to the two classes were strikingly similar. ..Apparently high ratings cannot be 'bought'..."

Comment: It is unfortunate that the authors did not take the very simple step of placing opposite each other the section grade point averages given by low-rating and high-rating classes. Instead they elected to give only the chi squares of difference in grades in the two classes. This is of course not very helpful to the reader since it deprives him of the opportunity of seeing whether the majority of the higher rated classes got somewhat higher grades, and it also does not allow him to see if the highest ratings, those in the 8th, 9th, and 10th deciles, were accompanied by high section grades.

If either of these situations existed, one might be tempted to take the Voeks and French studies and put them under the column in Table D that shows at least moderate associations between grades and ratings.

Professor Eble often points to the University of Washington as the model of good evaluation practices. In "What Are We Afraid Of?" he criticizes the "abuse of research" shown by Miriam and Burton Rodin in their article in Science that suggested students rate highest those teachers from whom they learn the least. Professor Eble used the following terms:

> The authors (Rodins) ommission of relevant research is curious. Though Virginia Voeks article "Publications and Teaching Effectiveness" is cited, a more relevant article by

> Voeks and G.M. French," Are Student-Ratings of Teachers Affected by Grades," with conclusions again the opposite of the authors', is not. Perhaps this is because Voeks' work is based on careful study of data amassed at the University of Washington, where almost 50 years of experience with student evaluations supports the conclusion that student evaluations do correlate with teaching effectiveness.

It seems to this reviewer that the data amassed at the University of Washington may have been somewhat distorted by the student rating scale used to amass it. It is the one encountered in the study of the large literature of student evaluation that is most curiously lacking in parallel structure. The first four rating categories: 1) very superior, 2) superior, 3) competent, 4) only fair, are standard enough, but the fifth category "of less value to me that the others" suddenly invites the student to switch from an evaluation of the instructor to an evaluation of the course. Even though Professor Eble went out of his way in his AAUP-AAC-Carnegie supported study, The Recognition and Evaluation of Teaching, to praise it as a model for other colleges to copy, it is difficult to see how valid rating data could be collected from it.

The Voeks and French studies have been of major importance to the literature of student evaluation of faculty. They are quoted in almost all the important review literature of the past dozen years. Their publication in 1960 negated the effects of the Anikeef (1953-3) and Weaver (1960) studies that had shown positive grade rating correlations. In the opinion of this reviewer the Voeks and French studies had the following faults:

1. The instrument used to collect the data was questionable.
2. The method of collecting the ratings invited bias.
3. The data collected was not displayed in the most natural way.
4. The conclusions reached did not follow naturally from the data that was displayed.

1953-1: A. W. BENDIG. "The Relation of Level of Course Achievement to Students' Instructor and Course Rating in Introductory Psychology". Educational and Psychological Measurement, 13, 1953.

Bendig studied 5 introductory psychology courses (132 students total) at the University of Pittsburgh in Spring 1951. He found positive

correlations of .14 to .28 between grades and ratings. He concludes "Student achievement does affect the rating, but not to a degree that invalidates continued use of the scales."

1953-2: A. W. BENDIG. "Student Achievement in Introductory Psychology And Student Rating of the Competence and Empathy of Their Instructors." Journal of Psychology, 36, 1953.

In fall 1951, Bendig again studied 5 sections of introductory psychology (121 students). Grades were apparently based entirely on objective achievement tests. He found strong negative correlations (figure not given) between grades and rating. He cancels his spring findings as follows "....the previously reported strong positive correlation between student achievement and summed ratings on the P.R.S.T. scales was a function of the factorial complexity of the scales." Bendig gives a possible explanation for the negative correlation, "Students of high overall ability may be more aware of inadequacies in the teaching of their instructors and to judge them more critically."

Comment: The presence of one unpopular high grader in a sample of five or a highly structured course, earning the contempt of abler students, could both produce the results. Bendig's high negative correlation and Heilman and Armentrout's -.042 are the only studies of the 41 to show negative correlations between grades and ratings. If no correlation existed between grades and ratings, approximately 20 studies could be expected to show negative results.

1954: KENNETH E. CLARK AND R. J. KELLER: "Student Rating of College Teaching." in R. E. Eckert and R. J. Keller (eds.) A University Looks at Its Program. University of Minnesota Press. 1954.

A total of 15,000 ratings by students in 380 classes in the University of Minnesota College of Science Literature and the Arts were collected in a voluntary program in 1949. Though the authors supply no specific data, they report they found little relationship between the students overall grade point average as he reported it on the rating form and the teacher ratings. In fact, students with grade point averages below "C" tended to rate teachers somewhat higher in general teaching ability than other students. Only a few items such as quality of exams, ability of teachers to adjust to the level of the students and willingness to recommend the course to a friend were found to be related to ratings.

Comment: The interesting tendency of truly marginal D and F students to rate their teachers higher than C students has been observed in several other studies. The Clark and Keller study, concerns itself only with the grade point average that the student brings to class. It is not a study of grades earned or expected within a class. Most grade point average and placement test studies tend to agree that basic student ability is only marginally related to teacher ratings. Clark and Keller's study, of course, says nothing about the relationship between ratings and the grades the students were expecting from the instructor they were rating.

1962: C. M. GARVERICK AND H. D. CARTER: "Instructor Ratings and Expected Grades." California Journal of Educational Research, 13, 1962.

Sample: 164 students of one instructor in an introductory psychology course at Berkeley in two semesters. Findings: The grades the student expected and the grades the student thought he deserved had little relationship to teacher rating. The correlation was only .079.

Comment: A one teacher study of this type proves little. See the comments under 1936-2 (Blum).

1966-1: C. L. OVERTURF AND E. C. PRICE: "Student Rating of Faculty at St. John's River Junior College With Addendum for Albany Junior College." 1966. ERIC Document EDO 13066.

A total of 10,000 ratings were taken college wide in 1964-65. The ratings were compulsory, the Dean of the college and the instructor waiting outside the door while the students completed the forms. The results were apparently used for merit pay and other personnel purposes. Teachers were ranked 1 to 91 according to their evaluations. Although the highest ranking instructor gave 72% A's and B's and the lowest 7.2% A's and B's, the authors report that when they ranked the 91 pairs (mean gpa given by teachers and mean ratings given by students) and applied Spearman's rank order equation to the two lists, they found a correlation of only .17, significant at the 10% level but not at the .05. Following the common statistical custom of not finding an association unless there is 95% certainty that the results could not have come about by accident,

they state, "The statistical evidence does not support the conclusion that instructors awarding higher marks should expect a higher rating from his students."

Comment: The statistical evidence from St. John's River is unconvincing. Anikeef and several others have used Spearman's formula to find correlations between grades and ratings, correlations which incidentally turned out to be much higher than Overturf and Price found, but they worked with lesser numbers. The Spearman formula is believed to be accurate enough for most purposes when 15 to 30 pairs are being correlated, but it does not seem reasonable to expect it to handle 91 pairs. Overturf and Price were surely working with very large squares of difference in rank and with a number of ties needing correction. A. C. Crocker in Statistics for the Teacher, 1971, says on page 58, "A simple method of calculating a correlation is the Spearman rank order correlation. This is useful for classes of children (or any set of scores) up to a maximum of thirty scores in each set. Beyond thirty the results tend to be unreliable."

1969-2: BERNARD CAFFREY: "Lack of Bias in Student Evaluation of Teachers." Proceedings of the 77th Annual Convention, American Psychological Association, 1969, Vol. 4.

Caffrey studied 131 students in three sections taught by three different instructors at Clemson University. The subject matter taught or methods of grading are not discussed. He found that only 6 of the 46 items on the rating scale correlated beyond the .01 level of significance with course expected grades and only two correlated at that level with grade point average. The six positive correlations with expected grade ranged from .32 for the students overall rating of the course to .23 for the instructors ability to explain clearly. The author judged the effect of grades on ratings to be of little importance.

1971-3: MILTON HILDEBRAND, ROBERT C. WILSON AND EVELYN R. DIENST: Evaluating University Teaching. Center of Research and Development in Higher Education: University of California, 1971.

The authors undertook a study at the University of California at Davis designed to develop a rating system. As part of the study they took 1015 student ratings. The method of collecting data is unclear and no specific

data is listed. They do however state that they found small positive correlations with grades which were significant at just beyond the .01 level. Their findings they believe are consistent with previous research: for they comment: "Cohen and Brawer (1969) reported similar results. Other studies have reported a relationship between expected grades and ratings of teachers (Stewart and Malpass, 1966; Weaver, 1960), a relationship only at lower class levels (Anikeef 1953), and no relationship (Kent, 1967, Voeks and French, 1960). These contradictions seem consistent with the presence of a definite but trifling correlation."

Comment: Hildebrand and his associates seem to be mistaken on nearly all counts when they check the believability of their own findings by reference to past research. The original research study in Cohen and Brawer (1969) seems to say nothing about grades. Instead Cohen and Brawer refer to the doubtful Overturf and Price study (1966-1 above). A careful reading of Anikeef (1953-3) shows he did find associations between grades and ratings at the upper class level -- .43 to be exact. Kent is a secondary source. Voeks and French (1952-60) did of course report no correlation.

1972-3: ALLEN C. KELLEY: "Uses and Abuses of Course Evaluations as Measures of Educational Output." Journal of Economic Education, 4, No. 1, 1972

Sample: 258 students in two lecture sections in economics at the University of Wisconsin, Madison. Both sections were taught by the same professor. He was aided by 7 graduate assistants who met discussions sections once a week. The ratings were taken after the first midterm exam and before the second. Controls: Though ratings could not be anonymous, the students were assured that their identities would not be revealed. The senior professor left the room when ratings were taken. Controls not discussed are: first, the nature of the midterm exam, whether it was objective or essay and whether graded by computer, assistants, or senior professors: and second, whether or not the students believed the results would be used for personnel purposes. Findings: By constructing two simulated statistical models, projecting what would have happened if conditions in the course had been different, Kelley demonstrates that if students had received only A's and B's for the midterm exam the actual rating of the senior professor would have risen from 3.784 to only 3.860. Thus he finds

that the impact of increased expected grades, though statistically significant, was very minor. The teaching assistants as a group were found to produce a negative effect on the senior professors rating. The negative effect was caused largely by TA#5. If his students had been enrolled in the classes of TA#2 and TA#4, the senior professor's ratings, according to Kelley, would have been .28 higher.

Comment: No comment.

The Relevance of Past Research to the Harper English Findings

The evidence indicates that the widely-held belief that grades and ratings are unrelated is a myth. Further, it indicates that the myth seems to have been spread by those who have a vested interest in promoting it. The body of empirical research that supposedly underlies the no-relationship generalization turns out to be without real substance when one makes an effort to look at all the evidence, not just at selected studies or parts of selected studies. If a convincing body of evidence exists to support the generalization it has evidently not been published.

The "classic research" as McKeachie called it, of H. H. Remmers and his students at Purdue turns out to be: (1) an examination of four college instructors (subjects and methods of grading unstated) and seven high school practice teachers; (2) 37 graduate assistants in a chemistry course where the senior professor was not evaluated, where the assistants for the most part did not plan to become teachers and had little to say in assigning grades; (3) a study of an unstated number of Purdue instructors teaching unstated subjects, the study design guaranteeing that differences in student reaction to hard and easy teachers would be concealed; (4) an instruction manual reporting the findings of the research and in doing so turning one college into 10, an unstated number of instructors into 460, 3786 students into 26,014, and "virtually ignorable" differences into totally ignored differences.

The "careful study of data," as Professor Eble describes it, at the University of Washington, which next to the Purdue studies did the most to promote the no-relationship generalization, is revealed upon examination to be somewhat less careful than one might wish. And the same is true of the project at St. John's River Community College, a project much publicized among two year colleges, where the dean and the teacher stood together outside the classroom door, waiting for the ratings that would rank the teachers in order 1 through 91. Of the 11 remaining no-bias studies, three are one-teacher in-class projects; one is a three teacher study finding correlations of up to .32; two are conflicting five teacher studies; and one is a grade point average study (Clark & Keller). One of the remaining studies (Starrack, 1934) reports a positive correlation of .15, but give no details as to how the figure was

reached. Another (Hudelson, 1951) seems to prove that grades are quite important when a correlation of .19 exists.

The case for the no-bias position rests largely on two studies, conducted 35 years apart. These are the Heilman and Armentrout study of 1936 (correlation -.042) and Hildebrand, 1971 (correlation about .09), and the former is tainted by the size of the standard deviations and the latter by the lack of specific details about how the grade-rating correlations were drawn, the type of classes used as samples, and the like. In any event one must view them in conjunction with a large number of studies, many quite persuasive, that show otherwise. If we were to draw a frequency curve of all the 28 published grade-rating studies made since 1953, including all six studies in Column 3, translating all findings into correlation coefficients, the range would run from just under +.10 to +.90 with a fairly even distribution between. There seems to be a tendency for the correlations to be higher when the ratings are compulsory and tied in with a merit pay or faculty promotion system. They also seemed to be higher when grading is subjective and when the teachers being rated are teaching multi-sectioned courses.

The Harper English study findings of 1973-74 and 1974-75 therefore do not seem to be atypical. Rather they seem to fall easily into the patterns established by prior research. There is little doubt that a strong relationship between grades and rating exists in the English Department at Harper. It is doubtful that as high a correlation exists in other departments and divisions of the college, but it would be most surprising considering the history of the research to find any large transfer course subject area where it did not exist in some form.

It is customary among some statisticians to assert that correlation of less than a .25 as virtually meaningless and those of under .50 as indicating something of no great importance even when constant replication of results indicate that association exists beyond reasonable doubt. But when people are being ranked on a scale the assertion would seem to be open to question. Other things being equal the one with only a slight advantage will be ranked ahead. Hudelson's correlation of +.19 and Bassin's (1974-2) correlation of +.10 with their corresponding shifts in percentile ranks illustrate this.

When a limited number of promotions are being competed for, even small correlations become meaningful. The spread between the 4th and 10th teacher rank deciles in most rating scales usually does not exceed one half of a rating level. Harper's is no exception.

The Harper English study shows that good ratings and moderate grades are not incompatible. It also shows that giving high grades does not of itself guarantee high ratings, but it does show, beyond doubt, that on journeys to the high country - the 8th, 9th and 10th deciles where the "outstanding" English teachers are - high grades seem to be essential. A recent Harper College committee report suggests that the term "outstanding" be reserved for those teachers who had scored 5.50 or above on the 6 point Harper rating scales. Only four English sections in the fall 1974-75 term reached that level. The average expected grade in the four courses was 3.27, the average final grade 3.07. In expected grade means these sections rank 1st, 7th, 8th and 11th among the 35 in in the study. In final grade means they rank 1st, 2nd, 4th and 8 th. The grading style indexes of the four teachers ranked 1st, 2nd, 3rd and 6th among the 16 teachers in the department.

At the bottom of Table B are the four lowest rated sections in the department. They are in first decile college wide as well. These sections ranked 20th, 26th, 34th and 35th in expected grade, 19th, 29th, 32nd and 35th in final grade. Their teachers ranked 12th 13th, 15th and 16th in their grading style indexes.

These rankings are quite consistent with the findings of numerous empirical studies during the past 20 years.

High Grades, High Ratings and Student Learning

Once the correlation between student grades and teacher rating is demonstrated it becomes necessary to prove that the grades are earned by the student rather than given freely by the teacher. Otherwise, the belief in the validity of ratings must collapse. Otherwise, no teacher can be certain of the degree to which student opinion of his knowledge and technique is colored by one aspect of his teaching - grading style.

For 30 years researchers have been trying with little success to prove a connection between learning and ratings. There is a sizeable literature on the subject. The studies are no more convincing than those that tried to prove that there was no correlation between grades and ratings. There is little need to summarize here all the studies in the literature that address this problem. Neither Professor Eble or anyone else seems to have claimed that a connection has been convincingly demonstrated, at least with rating scales that give the student an opportunity to state his preferences.

Several years ago a research study measured student writing improvement in English 101 sections at Harper. Nine of the 16 teachers included in the 1973-74 grade-rating studies participated. Four of the nine were high graders, five low. Their grading styles have not changed relative to each other since, though there has been an updrift of departmental grades as a whole. In the student achievement study, numerous gradings were made of paired start-of-semester and end-of-semester papers of 600 students.

The study did not attempt at that time to examine the relationships between grades and ratings. Its focus was only on an attempt to determine if there had been student achievement during the semester, and how much. It was found that in the most successful sections 40% to 50% of the students were writing better at the end than at the beginning. It is now possible to go back to that study to see if there was a relationship between grading styles and student achievement. <u>There was no grading-achievement relationship at all</u>. Of the three whose

classes showed the most improvement, one is a high grader who scored high in the 1973-75 student ratings. The other two had then and have now the lowest grading style indexes in the department. Both were at or near the bottom of the evaluation rankings in 1973-74 and at the lowest deciles departmentally and institutionally in 1974-75. Neither placed a class as high as 5.00 in the ratings.

What of other, more formal, studies? Remmers and Elliott attempted to show a connection in 1949 and 1950 in the sections of the chemistry assistants, but the correlations were difficult to pin down and shifting. Russell and Bendig, following the Remmer and Elliott example, in 1953, divided psychology students into a plus group consisting of those whose final grades were higher than pretests predicted, and a minus group whose grades were lower than predicted. They found, as had Remmers and Elliott before them, that slightly higher ratings came from the plus group, but Bendig, working alone, had found in study 1952-2 that the students who got the highest grades on the final exam appreciated their teachers least.

Recently more interesting work has been done. Peter Frey of Northwestern University, writing in the October, 1973 edition of Science, tells of a study of 13 calculus classes which showed correlations of up to +.90 between student mean section final grades and mean section teacher ratings. It was a study that could have been used in Table D to show a relationship between grades and ratings, but was rejected because its only focus was on student achievement. Grading in the sections was by a departmentally prepared final with a departmental curve. The individual teachers could neither be praised or blamed for being hard or easy graders. Frey argues that the high positive correlations between grades and ratings come about because his rating form identifies superior teaching. There is a weakness in his study in that the ratings were taken by mail after the student knew his final grades. They were not anonymous and students with low grades did not respond in the same proportion as students with high grades. So there is obviously no way of determining whether the ratings resulted from the student's discovery that he had made good grades or from his appreciation of good teaching. An influx of ratings from students with low grades might have driven the ratings down.

There is however good reason to believe that he has demonstrated a relationship between teacher ratings and student learning. His success seems to lie in his rating form, which is radically different from those in common use. Frey is not in favor of the global rating forms that measure student preference, such as the Purdue, CEQ and SIR types that have been popular since 1930. He does not use such questions as "Should this instructor be retained if suitable replacements are available?" that Remmers had on the form he used to rate the chemistry assistants or the "excellent" to "very poor" instructor ratings on the CEQ or the percentile rating of instructors on the SIR. Instead, Frey's key question asks the student to tell how much work he was required to do - not whether he liked doing the work or whether he liked the teacher who assigned it to him, but simply how much there was. This question combines with another on clarity of the instructor's presentation to produce, according to Frey, positive correlations in the neighborhood of .90 with student achievement as measured by final exams. Frey explains it bluntly: lack of clarity in the teachers presentation can be compensated for by a heavier student work load; if there is a heavy work load, explanations need not be so clear.

If the Frey scale were used instead of the student preference type now used, the two low-rated Harper English teachers mentioned above, both of whom assign large amounts of work, might be expected to rise rapidly in the ratings, even to the point where they might expect to be considered for promotion. But there is little chance that quantitative scales of the Frey type could be adopted in teacher merit systems. Setting teachers in competition with each other to see how much work they could assign would surely cause enrollments to decline rapidly.

A global student preference rating scale was used by Arthur Sullivan and Graham R. Skanes in 1972 at Memorial University of Newfoundland when they found a correlation between final exam grades and teacher ratings of .35 in 130 sections. The final exams, counting 50% of final grade, were departmentally prepared and graded. The sections, all of them in the sciences, math and psychology, worked from structured common syllabi. Therefore, it would seem that the effect of individual grading styles was partially cancelled. The study unfortunately suffers from two weaknesses that make it difficult to point to the study as proof that student achievement and grades are related. First, the authors did not

undertake the difficult and doubtful job of "regressing" the final exam score to compensate for differences in initial abilities in the different sections; second, in a follow-up study of 24 psychology teachers, the group that was found to have produced the strongest second year psychology students was a small group of low-rated teachers.

The best known rating-achievement study is that of Miriam and Burton Rodin published in Science in September, 1972. They used a global preference rating scale to find high negative correlations between the amount students learned and their rating of teaching assistants. Like the Remmers and Elliott studies of 1948 and 1950, the assistants had little to do with assigning student grades. The highly structured organization of the course, the exams and the methods of grading were the creations of the senior lecturer, who taught the class three of the five sessions each week. He was not rated by the students. The Rodins concluded that some of the assistants forced their students to work harder than others, and received low ratings as a result, even though their students scored higher on exams. Hence, the negative grade-rating correlations. The study has been vigorously attacked by numerous supporters of student ratings, among them Professor Eble, who says:

> Ignorance continues to appear. Last fall I was invited to Virginia Commonwealth University to discuss evaluation of teaching. Among the first things that confronted me when I arrived was an article just printed in Science called 'Student Evaluation of Teachers.' The subtitle made the claim: 'Students rate most highly instructors from whom they learn the least.' I spent a good part of the afternoon on the health science campus pointing out that the research that supposedly supports this claim was pretty shabby even by a humanist's standards.

Professor Eble is on sound grounds, though "ignorance" is perhaps not the term to apply to the study. The Rodins after all made a number of improvements in the design that Remmers and Elliott used in their work with teaching assistants at Purdue, the studies that succeeded in persuading large numbers of people that rating scales were valid. The conclusions the Rodins draw are supported by their data far better than those drawn by Voeks and French, whose work Professor Eble often praises. However, the sample was inadequate and the Rodin's conclusions should be approached cautiously.

Another atypical sample is found in the less publicized report by Richard Turner and Robert Thompson, (ERIC ED0900826) who report that a study of graduate students teaching 16 sections of French in 1972-72 and 24 sections in 1972-73 found substantial replicated negative correlations between student performance on exams and the students rating of the performance of the graduate assistants.

The evidence of learning-rating associations is weak. It seems unlikely that convincing positive correlations between the amount the student learns and the rating of instructors will be demonstrated soon if student preference-type global ratings continue in use as in the past.

The evidence indicates that the problem in getting the correlations between grades and ratings is not caused by a lack of student appreciation of teachers who are skilful in furthering student learning. It shows that most students do want to learn and do appreciate teachers who know their subject and can explain it clearly. The problem seems to be that students also appreciate other things in addition to learning. Apparently there are students sitting in every class who need something else more than they need to learn the subject, and their presence distorts the class mean and confuses the learning-rating issue. The need for praise, for example, is very strong in some students, as it is in teachers; the need for grades in some others. Sometimes the need for grades seem to be so strong that it outweighs all other considerations. A student needing a "B" to get a scholarship, or to stay in school or to transfer to another school or to graduate might find a course a disaster if he gets a "C", even though he learned a great deal.

Discrepant Grades and Scale Validity

The mechanism by which the grade-rating bias may work has been described in two of the studies summarized earlier. These are studies 1972-1 (Bausell and Magoon) and 1972-5 (Holmes). Together the studies suggest that two types of discrepant grade expectancies are operating in the classroom. Holmes has shown that students who are receiving lower grades than they anticipated may react by rating a teacher down in almost all aspects of his teaching technique. This can be termed negative discrepant grade reaction. It was found that the drop between mean section expected grade and mean section final grade was almost twice as severe among low grading teachers as high grading in the Harper English study. The high graders at Harper gave final mean section grades only a quarter grade lower on the average than the students expected. Low graders gave final grades averaging a half grade lower.

Bausell and Magoon in their study not only detected the negative discrepant grade reaction, but found a positive discrepant grade reaction as well. Students who were expecting higher grades than the grade point average they brought to the class tended to rate their teachers higher than expected.

The two types of grade discrepancies might, therefore, have influenced the Harper study results. The negative reaction could have occurred when the student found he was receiving lower grades than he had expected to receive, or suspected that his final grade would be lower. It could also occur when he found that he was receiving lower grades, or had to work harder for his grade, than his peers in other sections of the same course.

The present reviewer has seen evidence of the negative discrepant reaction as it occurred in an English 101 class during the 1973-74 fall semester. A high-grader was teaching an unusually weak section. On a Monday, a week before the end of the semester, he returned the last of a series of important tests to the class. Most of the students had done poorly. He administered the required faculty rating form immediately thereafter. When he examined his ratings he discovered he had received ratings much lower than he expected. On the following Wednesday,

he announced that the last test had been cancelled, and scheduled a new one. Two days later he gave a much easier test, and returned the papers on the following Monday. The average student had improved his grade one grade level. Upon taking another teacher rating immediately thereafter, he discovered that his ratings had within a week improved almost half of a rating level, enough for him to become a candidate for a 5% salary bonus then offered by the Board of Trustees to outstanding teachers. The Schuh and Crivelli study (1973-6) describes much the same student reaction.

The positive discrepant grade reaction, on the other hand, could occur when a student taking, for example, English 101, encounters a teacher who gives him higher grades than he received in high school English or praises his papers more than they have been praised before. It is possible that such a student would not only feel good about his teacher but might actually believe he had learned more than unprejudiced before and after testing could detect.

Both types of grade reactions would probably have their strongest effect in multi-sectioned, non-quantitative courses like English, where grading must be largely based on the subjective decisions of the teacher. The Holmes study suggests, however, that even when grading is done entirely through objective exams and the student can blame no one but himself, the disconfirming of grade expectations can have a strong effect on ratings.

Grade differences between those teachers with high ratings and those with low means that administrators or peer committees are asked to do an impossible job in interpreting preference rating scales. In looking at high ratings they must determine if praise or forgiveness in the classroom exceeded the bounds of intellectual honesty, knowing full well that positive reinforcement through grades may be the mark of a good teacher. In looking at low ratings they need to determine if strict adherence to traditional work load standards or to the college's official grading policy is a sign of bad teaching. The presence of a grade effect suggests that teacher evaluation systems based in whole or in part on student preference voting have always been invalid and may have lowered the quality of college teaching, not raised it, as supporters contend.

* * *

When I first started looking into the fringes of the literature about student evaluation of faculty a year and a half ago, I was a supporter of the use of student evaluations as an important part of a faculty evaluation system. I had used them in my own classes for almost 20 years and partly because of them had become a relaxed, permissive, high-grading teacher. I am now convinced that quantified student ratings of the preference type, even when used privately by the teacher for the avowed purpose of improving instruction and never shown to anyone else, have done more harm than good. The problem seems to be that the need of the students to be loved, praised and rewarded, and the need of the teacher to be loved, praised and rewarded, and the need of disciplines for their traditions, and the need of society for standards, do not quantify together in any rational way.

There are weaknesses in teaching that quantified scales help to correct, but there are strengths they tend to destroy, and on balance they seem to destroy more than they correct. One thing emerges clearly from a close study of the literature. The scales cannot discriminate between "good" teaching and "bad" teaching. A good reading test used in English classes cannot reliably discriminate between the 80th and 90th percentiles in student abilities, but it discriminates very well between the 10th and 90th percentile. Not so the teacher rating scales. There is little reason to believe that those English teachers who rank above the 90th percentile in cumulative student preference ratings are "better" teachers than those who rank below the 10th.

TABLE A

English Teacher Grading Styles

Based on Average Mean Section Grades

(1) FALL, 1973	(2) FALL, 1974*	(3) Grading Style Index		(4) % Of A's FALL 1973	(5) % Of A's Fall 1974
		High Graders			
3.14	3.16	(1)	3.15	31	29
3.02	2.75	(2)	2.92	39	26
3.01	2.84	(3)	2.88	29	22
2.92	2.72	(4)	2.82	23	18
2.63	2.78	(5)	2.70	38	44
2.74	2.64	(6)	2.69	13	27
2.66	2.66	(7)	2.66	30	31
2.40	2.76	(8)	2.58	24	29
		Low Graders			
2.40	2.56	(9)	2.48	7	12
2.51	2.26	(10)	2.39	19	8
2.27	2.27	(11)	2.27	10	13
Unavailable	2.13	(12)	2.13	X	21
1.92	2.26	(13)	2.09	9	14
2.08	2.05	(14)	2.07	6	9
2.06	2.06	(15)	2.06	7	12
2.17	1.71	(16)	1.94	16	8
2.53	2.48*		2.49		

* 1973 grades: A=4, B=3, C=2, D=1, F=0

1974 grades: A=4, B=3, C=2, D=1, F=0, N=0

The "N" grade had the effect of lowering teacher grade point averages an unknown amount, since some teachers used it to replace F's, incompletes and unofficial withdrawals. The latter two were not computed in the 1973 averages.

TABLE B

Teacher Evaluation Means in 35

English Sections

(6 Point Scale: 1VP, 2P, 3F, 4G, 5VG, 6EX)

Section Scores (1) of High Graders	Section Scores (1) of Low Graders
5.81*	
5.73*	
5.62*	
5.57 (2)	
5.47*	
	5.44
5.40	
5.38	
5.38	
5.33 -Median	
5.31-	
	5.29
	5.29
	5.27
5.22 Mean	
	5.18*
5.17	
	5.15*
	5.15*
5.14*	
	5.08*
5.00*	5.00
	5.00 Median
4.89	
4.79	
	4.71* Mean 4.78
	4.63
4.56	
	4.27
4.21	
	4.06*
	4.06
	3.89* (2)
	3.75

*Courses marked with asterisks are English 101 required composition courses. Those not marked are literature courses or other electives.

(1) Based on two year cumulative grading style index--average grades given to all students in two semesters a year apart. The grading style index range of the 8 high graders was 2.58 to 3.15 on a 4 point grading scale. The index for the 8 low grades was 1.94 to 2.48.

(2) Average grading style index of 4 outstanding sections, 2.91; of 4 worst sections, 2.06.

TABLE C

Rank Orders From the Section Giving Each of 16 Teachers His or Her Highest Rating in Fall, 1974

(1) Student Rating of Instructor: Section Mean	(2) Rating Dept. Rank	(3) Teacher Grading Style Index	(4) Index Dept. Rank	(5) Section Final Grade Mean: (A-N)	(6) Final Grade Dept. Rank
5.81	1	2.69	6	2.73	8
5.73	2	2.88	3	2.96	4
5.62	3	3.15	1	3.28	2
5.57	4	2.92	2	3.33	1
5.44	5	2.07	14	2.56	11
5.38	6	2.70	5	3.08	3
5.33	7	2.66	7	2.76	7
5.31	8	2.82	4	2.88	5
5.29	9	2.27	11	2.71	9
5.27	10	2.39	10	2.82	6
5.17	11	2.58	8	2.52	12
5.08	12	2.48	9	2.68	10
5.00	13	2.09	13	2.39	13
4.71	14	1.94	16	1.74	16
4.63	15	2.13	12	2.26	14
4.27	16	2.06	15	2.05	15

Spearman Correlation:

Column 2 with Column 4 = + .75
Column 2 with Column 6 = + .79

Both significant beyond the .01 level.

TABLE D

Correlations: Grades to Ratings

Comprehensive Listing of All Available 1930-74 Published Research	Sample	Association Judged Negligible	Association Judged Low to Moderate	Association Judged Marked or Important
(1)	(2)	(3)	(4)	(5)
1930 (Remmers)	11T,409R	+.07		
1934 (Starrak)	40,000R	+.15		
1936-1 (Heilman & Armentrout)	*46T	-.04		
1936-2 (Blum)	1T, 57R	Nil		
1949 (Remmers et al)	37T	+.13 to .35		
1950-1 (Elliott)	Unstated	+.03		
1950-2 (Elliott)	3786R	Unstated		
1951 (Hudelson)	192R	+.19		
1952-1 (Voeks & French)	34T	Nil		
1952-2 (Voeks & French)	20T	Nil		
1952-3 (Voeks & French)	16T	Nil		
1953-1 (Bendig)	5T	+.14 to .28		
1953-2 (Bendig)	5T	High Neg.		
1953-3 (Anikeef)	*19T			+.73 Merit
1954 (Clark & Keller)	15,000R	Unstated		
1960 (Weaver)	*12T, 699R			+.001 level
1962 (Garverick & Carter)	1T, 164R	+.08		
1964 (Enchandia)	16T		+Unstated	
1965-1 (Spencer & Dick)	600R		+Unstated	
1965-2 (Spencer & Dick)	*12 Sec.			+.85 to .91
1966-1 (Overturf & Price)	10,000R	+.17 Merit		
1966-2 (Stewart & Malpass)	*67T			+.001 level
1969-1 (Walker)	30T		+Unstated	
1969-2 (Caffrey)	3T, 131R	+.23 to .32		
1970 (Rubenstein & Mitchell)	*60 Sec.		+.09 to .44	
1971-1 (Holmes)	7 Sec.		+5% of Var.	
1971-2 (Wiegel et al.)	4T, 331R			+.01 level
1971-3 (Hildebrand)	1015R	+.09?		
1972-1 (Bausell & Magoon)	*12,000R			+.53 to .63
1972-2 (Nichols & Soper)	*339 Sec.			+.53 Merit
1972-3 (Kelley)	1T, 258R	+.02 of Var.		
1972-4 (Kennedy)	549R		+Unstated	
1972-5 (Holmes)	1T, 97R			+Various
1973-1 (Rosenshine et al.)	*1200 Sec.		+.09 to .27	
1973-2 (Perry & Baumann)	*900R			+.26 to .78
1973-3 (Centra & Linn)	300R		+Unstated	
1973-4 (Mirus)	*122 Sec.			+.85 Merit
1973-5 (Granzin & Painter)	*17 Sec.		+.14 to .21	
1973-6 (Schuh & Crivelli)	1T, 85R			+.001 level
1974-1 (Cornwell)	*70T		+11% of Var.	
1974-2 (Bassin)	*64T			+.10, 30% iles
Comparison				
1974 Harper (Powell)	18T			+.73
1975 Harper (Powell)	16T, 35 Sec.			+.43 to .79

T=Teachers R=Ratings Sec.=Sections Var.=Variance Level=Level of Confidence
Merit: The author indicates that ratings were used to decide pay, promotions, etc.

* An asterisk indicates the report met minimum research and research reporting requirements: typicality and size of sample, design, reporting of data, etc.

APPENDIX E

A Statistical Note

Two statistical symbols are used in this paper. The first, correlation coefficients, are estimates, reached through standardized algebraic formulas, of the degree of association between two or more sets of figures. They are stated in terms of departure from zero correlation (0.00). Perfect positive correlation is +1.00, perfect negative, -1.00. The + sign is usually omitted before positive correlations. A perfect 1.00 correlation would be produced if when grading on and being rated on a five point scale, the teacher received all fives from his A students, an average of four from his B students, an average of three from his C students, etc. A correlation of .50 would probably occur if a teacher received average ratings of 3.50 from C students, 4.00 from B students, 4.50 from A students. Average reatings of 4.10, 4.20, and 4.30 respectively would produce a correlation of .10. Equal negative correlations would occur if grades were inversely related to ratings. They would be of equal significance. Obviously, .79 is a strong showing, .10 a weak one.

The second type of statistical symbols, level of significance notations, are estimates of how likely it is that the results found occurred by chance. They are expressed as percentages. A .05 level of significance means that the results might occur by chance alone five times in a hundred; .01 indicates one chance in a hundred; .001, one chance in a thousand. It's important to remember that a high level of confidence does not necessarily mean that a high correlation is present if the sample is large. A small sample, on the other hand, must produce higher correlations before significance can be claimed. Several researchers, finding interesting correlations of .35 to .45 between grades and ratings have dismissed them as non-significant because they were working with small samples. A correlation of .49, for example, is needed to claim significance at the .05 level when one is working with a sample of 16 teachers. It's also important to keep in mind that some researchers, having found statistically significant but small correlations of .10, .20, and .30 have dismissed them as negligible, trifling, or slight. In doing so they are exercising statistical judgment, which may or may not be sound. Statistical practice is involved but not rigid statistical law.