12-1975

A Casual Analysis of the Relationship between Students' Expected Grades & Their Ratings of an Instructor

Dianne Willoughby
Western Kentucky University

Willoughby,

Dianne

1975

A CAUSAL ANALYSIS OF THE RELATIONSHIP BETWEEN STUDENTS' EXPECTED GRADES AND THEIR RATINGS OF AN INSTRUCTOR

A Thesis
Presented to
the Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
of the Requirements for the Degree
Master of Arts

by

Dianne Willoughby

December 1975

A CAUSAL ANALYSIS OF THE RELATIONSHIP BETWEEN
STUDENTS' EXPECTED GRADES AND THEIR
RATINGS OF AN INSTRUCTOR

Recommended 12 - 15 - 75
(Date)

Raymond M. Mendel
Director of Thesis

Sam S. McFarland

John R. Faine

Approved 12 - 16 - 75
(Date)

Elmer Gray
Dean of the Graduate College

ACKNOWLEDGMENTS

This study has not been solely the product of the writer. For this reason, acknowledgments are appropriate.

Thanks to my advisor, Dr. Ray Mendel, for initiating the study, and for consultation during selected phases of data collection and analysis. I appreciate the many readings you gave my rough drafts and your suggestions for their improvement.

Thanks also to Dr. Terry Dickinson for suggesting the statistical analysis--an obscure method for assessing causality which makes this more than just another study.

To Dr. John Faine, thanks for isolating the sample classes and giving me access to the data. Your comments and suggestions provided me with an interdisciplinary perspective on research which I hope to retain.

Additionally, thanks to Dr. Sam McFarland for serving as a member of my committee.

Special thanks to Carolyn Marks and Dr. Tom Madron for instruction, advice, and encouragement during the computer analysis phase of this study. You are both masters at keeping your head when all around you are losing theirs. The

technical advice I received from you runs a close second to the moral support you provided.

To my husband, Michael Bridgman, I want to give the final thanks. Being able to count on your constant encouragement and willingness to help meant more to me than I can express.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

A CAUSAL ANALYSIS OF THE RELATIONSHIP BETWEEN STUDENTS' EXPECTED GRADES AND THEIR RATINGS OF AN INSTRUCTOR

D. Willoughby August 1975 73 pages

Directed by: R. M. Mendel, J. Faine, and S. McFarland

Department of Psychology Western Kentucky Univerity

Is there a relationship between the grades students expect to receive in a course and the ratings they assign their course instructor? If a relationship does exist, do the students' grade expectations cause the ratings subsequently given the instructor? Data were collected at the beginning and end of a semester, and a cross-lagged panel correlational analysis was applied to two pairs of variables. The first pair of variables, a single-item assessment of instructor effectiveness and a single-item record of each student's expected grade, indicated a statistically significant relationship between expected grades and the measure of instructor performance. This relationship was stronger at the end of the semester than it was at the beginning, and cross-lagged correlations indicated that students' expected grades are causal contributors to the single-item overall instructor ratings. The second variable pair included the same measure of expected grade and a factor score measure of instructor performance. The cross-lagged data from this variable pair also showed a stronger grade-rating relationship at the end of the semester than at the beginning. However, the hypothesis that expected grades cause factor-score instructor ratings was not confirmed.

CHAPTER I

INTRODUCTION

In the past twenty-five years, student evaluations have become a widely used method of instructor performance appraisal (Mueller, 1951; Gustad, 1961; Bassin, 1974). Many studies have examined the properties of student evaluations in order to determine whether or not the information these evaluations provide is reliable, valid, and sufficiently unbiased (Costin, Greenough and Menges, 1971). This paper directs itself primarily to potential determinants of student evaluations. The main focus is on the relationship between expected grades and instructor rating. In order to justify the selection of this potential determinant for investigation, as well as to provide the reader with an overview of current research in the area, the literature reviewed in the first half of the following section is divided into three parts.

Part 1 contains a summary of studies designed to determine the reliability of student evaluations. Although the results of studies reported in this section are determined more by the skill with which the evaluation instrument is developed than by the characteristics of the student raters, consistency of results across instruments can be demonstrated. Part 2 contains studies which address the issue of validity. The construct validity of student evaluations can be determined by the extent to which these evaluations measure the construct "teacher performance." A method frequently used to test the construct validity of student evaluations involves comparing these evaluations with other accepted measures of teacher performance to determine the extent to which the two agree. Part 3 discusses potential sources of bias and reviews research related to the causal relationship between expected grades and instructor ratings.

Literature Review

Part 1: Reliability studies

Opponents of student evaluations claim students are not well qualified as raters on the grounds that their ratings are not stable over time. However, studies designed to examine the stability of student ratings typically refute claims of unreliability (Kohlan, 1973; Centra, 1973). Guthrie (1954) found correlations of .87 and .89 between student rankings of the quality of their teachers from one year to the next. He also reported that such judgments were more stable than faculty judgments of teaching quality in the same instructors.

Lovell and Haner (1955) computed split-half reliability on student ratings obtained using a forced choice rating scale and reported a mean even-odd item correlation (after Spearman-Brown correction) of .88. In another study, Voeks and French (1960) reported high inter-rater agreement on student ratings of instructors obtained at the same point in time (.94) and on ratings made two years apart by different groups of students (.87). Although each of these studies employed different evaluation instruments, the consistency with which reliable results were obtained suggests that when care is taken in the development of the instrument to be used for faculty evaluation, student raters can provide consistent ratings (Hall, 1965; McKeachie, 1969).

Part 2: Validity studies

Another criticism of student ratings involves the ability of students to accurately judge their instructors' performance. The accuracy with which such judgments are made is an indication of the construct validity of student ratings. Typically, convergent validity, which involves the comparison of the method in question with other proven methods of evaluation, is employed to compare students' ratings with established measures of instructor effectiveness.

Guthrie (1949, 1954) compared student ratings with ratings made by an instructor's peers and reported correlations ranging from .30 to .53. Using the same evaluation forms, and virtually the same method, Webb and Nolan (1955) obtained student evaluations and peer ratings for a group of university instructors. In addition, they also collected instructor self-ratings for the same group of instructors. They reported a significant correlation between instructor self-ratings and student ratings (.62), but no relationship was obtained between ratings assigned by students and those assigned by peers. These apparently conflicting results cannot be easily explained. Both studies employed the same evaluation form and, except for the addition of self-ratings by Webb and Nolan, the same method.

McKeachie and Soloman (1958) compared student ratings with a more objective criterion of teacher performance. They reasoned that better teachers should generate interest in the subject area more effectively than teachers with less skill.

"Interest in the area" was defined as the percentage of each instructor's students who enrolled in advanced courses in the department after completing an introductory course. Instructor ratings were obtained during the last week of classes for each introductory course. Nonparametric correlations obtained between instructors' rank-order on the two measures ranged from -.47 to +.63. These results indicate that _some_ teachers who are highly rated by their students tend to generate student interest in the subject area while others do not. McKeachie and Soloman (1958) suggest that a study designed to identify moderator variables, which might influence the obtained relationship (such as student's ability or course difficulty), would be helpful in explaining these results.

A third question relating to the construct validity of student ratings concerns the students' ability to accurately assess the long-term value of their courses. Everyone has, at one time, heard a former student remark, "I just hated Professor Jones, but now I realize how much I learned in her class." Remmers and Drucker (1950) designed a study to assess the truth of this statement. They sought to determine if the students' ratings and ratings assigned by alumni ten years after graduation were significantly different. Questionnaires were mailed to all alumni of a small university who graduated in a given year requesting each respondent to identify the "best" and "worst" teacher in the department of the respondent's major, and in the university as a whole. The university's instructors were assigned a rank based on the total number of "best" and "worst" replies assigned to each by the respondents. When this rank ordering of instructors was compared (via nonparametric correlation) with the rank order assigned the same instructors by students currently enrolled in the university, the result was a positive correlation of moderate strength (.40 to .68). Remmers and Drucker (1950) concluded that students are able to assess the value of a course in much the same way as alumni who have been out of college for ten years.

These studies bring up another question regarding the use of student evaluations. If student ratings generally are significantly correlated with other methods of instructor performance appraisal, why is it necessary to administer them? Costin, Greenough and Menges (1971) reviewed a group of studies designed to provide information concerning the construct validity of this method of appraisal. They summarized their report by favoring the continued and expanded use of student evaluations. They stated that although positive correlations are typically obtained between student ratings and other measures of instructor performance, these correlations are such that student ratings can be shown to provide unique variance not accounted for by any other method of performance appraisal.

McKeachie (1969) agrees with the conclusion drawn by Costin et al. (1971). He suggests that since students have a personal investment in the quality of the instruction they receive, and their ratings are based on an almost limitless opportunity for observation, student ratings can become a valuable source of information for instructors, students, and administrators alike. Not all investigators are this supportive, however. Many studies have investigated the influence of rating bias in student evaluations.

Part 3: Specific sources of rating bias

Heilman and Armentrout (1936) addressed the issue of rating bias in an extensive investigation of potential determinants of student ratings. In the course of their research, they identified several variables which they believed could influence the outcome of a student evaluation. Among these were several variables of "legitimate" influence, such as "instructor's training," and "previous teaching experience," and several other variables which can best be described as sources of bias. Some of the potential bias producing variables were: difficulty of course subject matter, class size, instructor's personal characteristics such as sex and temperament, whether the course was required or elected, maturity or grade level of the raters, and instructor's grading leniency. Although none of the variables investigated by Heilman and Armentrout (1936) were found to be significantly related to the results of student evaluations, the issue was not resolved to the satisfaction of all potential users, and investigation continues.

Of the sources of bias identified by Heilman and Armentrout (1936), grading leniency has emerged as a popularly recurring topic of investigation (Heilman and Armentrout notwithstanding), possibly because its influence is difficult to dismiss in light of inconsistent statistical evidence. This inconsistent evidence is partially due to differences in the evaluation forms used in each study, and the times during the semester when the ratings were made. Although published studies uniformly name the evaluation instrument which was employed, often information regarding the time of the instrument's administration is not provided.

Another characteristic which contributes to the confusion concerns the way each investigator operationalizes the variables chosen for examination. "Instructor rating" has been variously defined as a single-global assessment, total score on a questionnaire, scores on subscales of the total instrument, or any combination of the above. Similarly, "student grade" has been defined as the actual grade assigned each student in the course being rated, student-reports of the grades they expect to earn in the class, students' GPA, or "instructor's GPA." (The latter is determined by averaging all grades assigned by an instructor over a given period of time.) In the present discussion, studies which employ a common definition of each variable are grouped together.

Actual course grades and total instrument scores. In one of the first studies designed to focus on the relationship between student grades and instructor ratings, Remmers (1928, 1930) administered a student evaluation questionnaire to seventeen university classes. The administration took place after mid-term. Although students were not asked to sign their evaluation forms, when the ratings were completed names were read of students whose course grade to date was "above the class average"; and these students were told to make an X at the top of their answer sheets. A biserial correlation computed between ratings assigned by the "above average" group and ratings assigned by the remainder of the class was not statistically significant (Remmers, 1930).

Russell and Bendig (1954) and Remmers, Martin, and Elliott (1949) conducted similar investigations into the grade-rating relationship. The two studies were comparable in design, with one exception: Remmers et al. (1949) used class _means_ to obtain measures of both grades and ratings while Russell and Bendig (1954) focused on the _individual student_ rather than the class. Both studies began by developing a regression equation for predicting letter grades in introductory level courses from college entrance exam scores. The regression equations were then employed to identify classes (Remmers et al., 1949) or individuals (Russell and Bendig, 1954) who were earning course grades significantly above or below the course grades predicted from the equation. Both studies examined the college entrance exam score distributions for all subjects and concluded on the basis of chi-square analysis that no significant differences in ability existed across achievement groups. Ratings made by the overachieving classes or individuals were compared with ratings made by the underachieving classes or individuals. When the class mean was the primary unit of analysis, significant between group differences were obtained on twenty-three of the twenty-four subscales of the evaluation instrument (Remmers et al., 1949). When individuals rather than classes were examined, significant (between-subject) differences in rating occurred on the total evaluation score (Russell and Bendig, 1954). In both cases, overachievers gave more favorable ratings. The conclusion drawn in both reports suggests that the grade-rating relationship is more apparent when the students' achievement level is statistically controlled.

Actual grades and subscale scores. Several of the more recent studies have employed factor analysis to identify separate components of teaching behavior. Many evaluation instruments contain subscales of items which evaluate the course rather than the instructor. When results are reported in terms of a "total instrument score" it is not possible to determine how much of the reported relationship is due to course content measurements rather than instructor-related measurements. However, even when this distinction is made, results of the analysis remain inconclusive.

This characteristic of inconclusive results is emphasized in an examination of two articles written by the same author (Bendig, 1953a, 1953b). Bendig (1953a) collected instructor ratings from students in six sections of introductory psychology. These ratings were compiled separately on the basis of letter grades earned to date of rating by students in each section. Because the evaluation instrument employed in his study provides separate scores for "instructor rating" and "course rating," Bendig (1953a) obtained separate correlations between students' grades and each factor. A significant correlation (r = +.38) was reported between students' grades and their ratings of the course, while a nonsignificant correlation (r = +.14) was reported between students' grades and instructor ratings.

In his second article, Bendig (1953b) re-examined his data using factor analysis. Two factors emerged in place of overall "instructor rating," as defined in the first report (Bendig, 1953a). These factors were differentially related to the grades earned by the students who did the ratings. Factor II, Instructor Empathy, showed no relationship to the mean grade assigned by the instructor. Factor I, Instructional Competence, was negatively correlated (-.80) with the mean class grade assigned by the instructor, indicating that the students earning the highest grades gave the most critical ratings. Since both articles were based on the same data, the inconsistent results are especially difficult to explain. One plausible interpretation of these reports suggests that the definition of the variables examined in each case determined the outcome of the analysis. It is possible that the effects of the two factors, Instructor Empathy and Instructional Competence, "cancel each other" so that no relationship was obtained between the "instructor rating" which was a composite of the two, and student grade in the first study. When the variable "instructor rating" was more precisely defined, in terms of its principle factors, the significant relationship between one of these factors, Instructional Competence, and student grade was obtained. Through a process of successively redefining variables in terms of their principle components, it may be possible to identify factors which show consistent relationships with sources of bias such as students' grades.

Instructor GPA and total scores. Another method of operationalizing student grades involves the determination of the average grade assigned by each instructor over a period of two or more semesters. Investigators who define student grades in this way believe that "instructor's GPA" yields a measure of student grades that is not subject to the error involved in student self-report. Voeks and French (1960) approached the problem with this orientation. In their study, they computed rank order correlations for three hundred instructors who had been ranked on three criteria: (1) percentage A and B grades assigned in previous semesters; (2) percentage D and E grades assigned in previous semesters; and (3) total score on a student evaluation (where the professor scoring highest was ranked #1, etc.). Correlations between the two measures of grade and the total score were both nonsignificant. Voeks and French (1960) concluded that for classes on the whole, there was no relationship between an instructor's grading leniency and his standing on a student evaluation.

In order to determine whether results obtained from the total group of three hundred were consistent in special cases, the authors performed an analysis of the data collected from those instructors whose student ratings were considered extreme. Twenty instructors ranked either "best" or "worst" in each of ten departments were selected for the second phase of analysis. For each instructor, the mean of all grades assigned the previous term was computed, and comparisons were made between the ten "best" instructors' GPAs and the ten "worst" instructors' GPAs. Obtained differences, although not significant, were in the direction of a positive grade-rating bias. In light of these results, a third approach to the question was devised. Sixteen classes were selected in which repeated measures had shown a significant (three decile) increase in student ratings of the instructor. The GPAs of these instructors were then examined across semesters in order to determine whether a corresponding increase in the average grade assigned by the instructor could be identified. Based on the results of a chi-square analysis, no significant relationship was reported.

In this series of studies, Voeks and French (1960) have examined student ratings in order to determine what influence, if any, grading leniency has on their results. The three tests Voeks and French employed failed to provide evidence for a relationship between the two variables. Student ratings were examined for three hundred university faculty and no differences in ratings were obtained based on instructor GPA. When the "best" and "worst" instructor groups were identified on the basis of student ratings, no significant differences were obtained in the average instructor GPA across groups; and instructors whose student ratings showed improvement across semesters did not have a corresponding increase in instructor GPA. When reviewing these results, the variable definitions should be considered. As Bendig (1953a, b) has shown, "total score" on an evaluation instrument can reflect aspects of the course not directly related to the instructor's skill.

In a less elaborate study, Anikeeff (1953) compared student ratings of instructors with the instructor's GPA from the previous semester. Results, reported by grade level, show a strong correlation (.73) between an instructor's rating and the average grade he assigned underclassmen. The relationship is less strong for juniors and seniors (.43).

A recent investigation by Bassin (1974) employing similar definitions of instructor rating and instructor GPA reported low correlations between the two variables across sixty-three instructors. Instructor GPA ranged from 2.1 to 2.8 on a 4.0 scale. In this study, grading leniency accounted for less than 10 percent of the variance in obtained instructor ratings. This low relationship may be due, in part, to the lack of variance in instructor GPA. However, Bassin (1974) suggests that even this low correlation can have a dramatic influence on obtained ratings. An example was given in which increasing an instructor's GPA from 2.0 to 2.5, while holding all other variables statistically constant, resulted in a rating increase from the 30th to the 62nd percentile.

Few substantive conclusions can be drawn on the basis of studies reported in this section. Voeks and French (1960) reported no significant relationship between ratings and instructor GPA although they employed three separate designs in their investigation. Anikeeff (1953) reported moderate to high correlations, and Bassin (1974) reported correlations which, although quite low, were of practical significance. The differences in the strength of the relationships reported in these studies are probably a function of the evaluation instruments employed by each investigator. Because the total score on an evaluation frequently includes variance not directly attributable to the instructor, and evaluation instruments differ in the amount of variance each component is responsible for, it is possible that the investigators reviewed above were not examining exactly the same thing. For example, Anikeeff's (1953) instrument included fifteen teacher-behavior dimensions while Bassin's (1974) included only five. Their results might have been more consistent had their variable definitions agreed.

Expected grades and subscale scores. The studies reported thus far have focused on grades actually earned by the student or grades actually assigned by the instructor (Instructor GPA). Several authors have suggested that a more realistic source of potential rating bias is the grade the students _expect_ to make in the course at the time the ratings are taken. If student ratings are influenced by the rater's _expected grade_, studies which have examined _final course grades_ will reflect this bias only to the extent that student grade estimates accurately predict grade outcomes. If, as Schuh and Crivelli (1973) suggest, student ratings are a method of reprisal leveled at instructors on the basis of the grades the students are assigned, then it is the grade the student expects to make in the course at the time when the ratings are made (rather than the grade each student ultimately receives) which influences that student's rating of the instructor. For example, in a class where ratings are made two weeks before the end of a semester and only one test (a mid-term) has been given, student ratings may be highly related to the final grades the students expect to make based on their grades at mid-term. However, if scores on the final exam significantly change the students' standings in the course, the influence of this "adjusted" grade expectation (or final grade) can not possibly affect ratings which were made before the exam was given.

Statistical support for this reasoning is provided by Blum (1936). When students were asked to predict their final course grades, the accuracy of their predictions steadily increased as the semester progressed. Immediately before the final exam, predictions were 70 percent accurate. However, immediately after the final exam had been taken, the accuracy of student prediction increased to better than 85 percent (Blum, 1936).

In order to minimize the error associated with the discrepancy between expected grades and grades actually assigned, several researchers have selected "expected grade" as the variable to be examined. In one such study, Gaverick and Carter (1962) submitted instructor ratings to cluster analysis. Two clusters of items were derived from the total instrument and identified by the authors as "necessary and sufficient to account for the principle trends" among responses to the rating form. The first cluster contained items related to expected course grade, and the second cluster included items related to general instructor effectiveness. Because the rationale underlying the cluster-analysis technique involves identifying a "minimal number of _most nearly independent_ clusters which describe the general properties of the variable in question" (Tryon, 1958, p. 3), it is not surprising that the two clusters identified by Gaverick and Carter (1962) were not significantly correlated. Gaverick and Carter obtained their data from students in one large (n = 164) introductory class only a short time before the end of the semester.

Echandia (1964) also obtained data from one large introductory class by asking students to rate their instructor only weeks before the end of the semester. When the results of this evaluation were submitted to principle component factor analysis, three factors emerged. One factor contained items related to students' expected grades, a second factor described the instructor's classroom efficiency, and a third factor represented affective teacher-student relationships. Students' expected grades were significantly related to instructor efficiency factor scores; students who received higher grades had a significantly higher mean for total scores given the instructor on the efficiency factor. The correlation between these factors was also significant (r = .74). No significant relationships were obtained between the expected grade factor and the factor measuring affective interaction.

Although Gaverick and Carter (1962) and Echandia (1964) were addressing similar questions, their methods of analysis make comparisons of their results difficult. It is possible that had Gaverick and Carter refined their "general instructor effectiveness" cluster into its component parts, they might have identified two factors which parallel those described by Echandia (1964). In these studies, as well as those by Bendig (1953a, 1953b), it appears that relationships which exist between grades and specific components of instructor effectiveness can be obscured when a composite rating of instruction (rather than the component factors) is chosen for investigation.

Further evidence in support of this point is provided by Weaver (1960). In this study, instructor evaluation forms were administered to thirty-nine classes taught by twelve instructors. After the evaluations were collected, they were sorted into four groups on the basis of expected grade, and the mean ratings assigned the instructors by each group were compared. Weaver reported a significant relationship between expected grade and scores on items related to instructor competence. A similar relationship was not obtained between expected grades and items related to teacher personality, or between expected grade and the total evaluation score.

Caffrey (1969) reported comparable results based on an analysis of two factors identified as principle components of the student evaluation form used in his study. The factor which defined instructor competence was reported to be significantly related to expected grades, while the factor describing the affective abilities of instructors failed to reach significance.

Course grades and single-item global ratings. Several studies have examined the relationship between a global rating of overall instructor effectiveness and grades the raters expected or received. Kooker (1968) analyzed the responses students gave to an item designed to assess "overall" instructor effectiveness, in terms of the grades students had earned in the course. For each level, freshman through senior, a significant difference was obtained in overall instructor rating as it was assigned by groups of students earning different grades in the course. As a comment on his results, Kooker (1968) suggests that students may form impressions early in a course which consistently affect the students' responses to the course content, and subsequently, each student's performance. On the basis of available data, however, it cannot be determined what operates to produce the relationship Kooker obtained.

Treffinger and Feldhusen (1970) examined the results of a university-wide student evaluation of instruction. Using multivariate analysis, they identified a moderate relationship between expected grades and an overall rating of instructional effectiveness ($r = .39$). Data for this analysis were obtained from one large class ($n = 192$) based on ratings of the class instructor made at the end of a semester. No significant similarities were found, however, between these end-of-semester ratings and ratings the same students had made earlier in the semester regarding the "general quality of instruction in the university as a whole." Treffinger and Feldhusen (1970) suggest these findings represent complex patterns of interaction between the students' initial impressions of the course (based on hearsay, and the climate of the university or department), cognitive and affective characteristics of the class as a whole, and instructor performance. They conclude their remarks with the recommendation that an analysis of the extent to which each student's final rating of the instructor has changed from his initial impression of the instructor is necessary in order to more clearly describe the impact of their findings (Treffinger and Feldhusen, 1970).

In a less elaborate design, Schuh and Crivelli (1973) asked students in one large class to rate their instructor's performance by assigning him a letter grade. On the same card, the raters were asked to record the letter grade they expected to make in the class. The evaluation was conducted after the students had been told of their mid-term grades.

Ratings were collected and divided into four groups on the basis of the grade expected by the raters (A, B, C, D). The distributions of grades given the instructor by students in each group were then compared via analysis of variance and a significant rating difference was reported between groups. Schuh and Crivelli (1973) reflect upon their results in the following remarks:

> Clearly a small but significant portion of the variance in the student's ratings of faculty teaching effectiveness is a reflection of the student's mid-term grade. A suitable term for this source of bias in rating should imply the mirroring back to the supervisor of his evaluation of the subordinate's performance. Webster's Seventh New Collegiate Dictionary (1967) was consulted under terms with the connotation of reflecting blame. Animadversion was defined as a term implying criticism prompted by prejudice or ill will, hence, the adoption of the term animadversion to describe the error. (p. 259)

They suggest that the effect of animadversion error in student evaluation of teaching might be reduced if such evaluations were administered early in the semester before any examinations are given, thus depriving the rater of the "contamination information," that is, performance feedback in the form of a grade.

In summary, the literature concerning the relationship between student grades and instructor ratings is inconsistent and inconclusive. The inconsistency arises from the multitude of definitions employed by investigators, all of whom claim to be testing fundamentally the same relationship. Grades have been defined as actual course grade to date of rating, expected grade at time of rating, or average grade assigned by the instructor who is rated. Similarly, as stated before, instructor rating has been defined as total instrument score, score on a factor scale within the total instrument, or a single item score reflecting overall teaching effectiveness.

Even in those studies which employ similar definitions of each variable, the results are often conflicting due to differences in evaluation instruments, a restricted range of scores on one or both variables, or the omission of information in the report regarding when the evaluations were made or how the data were grouped (by class or by student) for analysis. Such studies may report significant grade-rating relationships and yet add little to our understanding of these relationships. Table 1 summarizes the literature reviewed to this point.

An unwarranted assumption often made by investigators who report a significant relationship between student ratings and grades concerns the issue of causality (Treffinger

TABLE 1

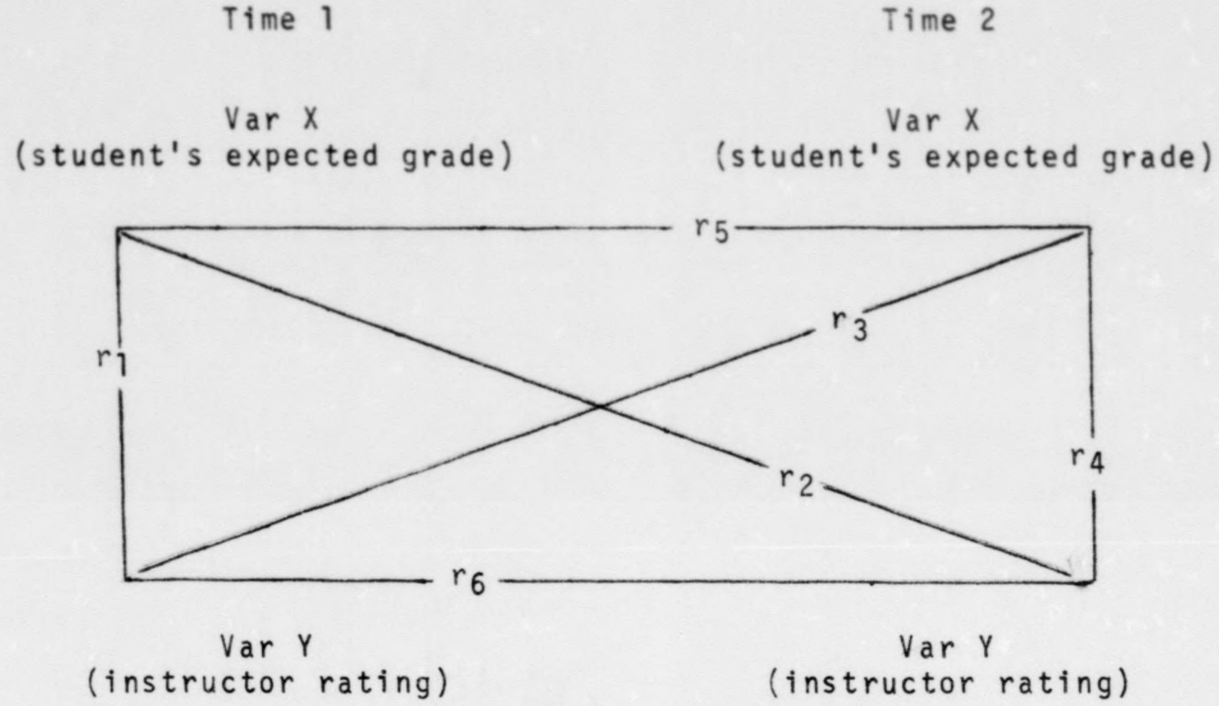
SUMMARY OF LITERATURE

Study	Grade Definition	Rating Definition	Significant
Remmers (1928, 1930)	Actual grades at midterm	Total instrument score	No
Russell & Bendig (1954)	Actual grades	Total instrument score	Yes
Remmers, Martin & Elliott (1949)	Actual grades	Subscale scores	Yes
Bendig (1953a)	Actual grades	Subscale scores	No
Bendig (1953b)	Actual grades	Subscale scores	Yes
Voeks & French (1960)	Instructor GPA	Total instrument score	No
Bassin (1974)	Instructor GPA	Total instrument score	No
Anikeeff (1953)	Instructor GPA	Total instrument score	Yes
Gaverick & Carter (1962)	Expected grade	Subscale scores	No
Echandia (1964)	Expected grade	Subscale scores	Yes
Weaver (1960)	Expected grade	Mean Subscale scores	Yes
Caffrey (1969)	Expected grade	Subscale scores	Yes
Kooker (1968)	Actual grade	Single-Item	Yes
Treffinger & Feldhusen (1970)	Expected grade	Single-Item	Yes
Schuh & Crivelli (1974)	Expected Grade at midterm	Single-Item	Yes

and Feldhusen, 1970; Schuh and Crivelli, 1973). The assumption, that the student's knowledge or estimate of his own course grade causes the ratings given to the course instructor, cannot be accepted without investigation. Although this assumption appeals to common logic, an equally plausible explanation might suggest that the students' initial impressions of their instructors cause them to perceive each course in a given way which ultimately results in the grades they earn. Or, that a third variable not formally considered may be causing the changes observed in both ratings and grades. Kooker (1968) has suggested that research aimed at determining what effect the purposeful alteration of students' perceptions of a course has on subsequent course achievement would help clarify the nature of the relationship. For obvious reasons, manipulation as suggested by Kooker is ethically (if not methodologically) questionable. Under the conditions typically found in college classrooms, the issue of causality can best be addressed using statistical techniques rather than experimental manipulation.

Part 4: Cross-lagged panel correlational model

Through a method discussed by Campbell and Stanley (1963), it is possible to test with some confidence the strength and direction of causal relationships by employing correlational analyses on repeated measures. The method, known as the cross-lagged panel correlation, is illustrated in Figure 1.



r_5 and r_6: Reliability measures of Var X and Var Y

$r_2 > (r_1 = r_4) > r_3 \longrightarrow$ Var X caused Var Y

$r_3 > (r_1 = r_4) > r_2 \longrightarrow$ Var Y caused Var X

Fig. 1. Basic cross-lagged correlation model

Each variable, variable X (student's expected grade) and variable Y (instructor rating), is measured at two points in time. The six possible intercorrelations are then computed and the resulting coefficients are examined for evidence of causality. Logically, if X causes Y, the correlation between X (time 1) and Y (time 2) should be greater than the correlation between Y (time 1) and X (time 2), because in order for X to cause Y, X must either precede or be concommitant with Y. If X follows Y, then Y cannot possibly be caused by X and the size of the correlation coefficients should be reversed. Referring to Figure 1, if students' expected grades cause instructor ratings, r_3 should be greater than r_2. This relationship would be reversed if instructor ratings cause student grade expectations. The correlations r_2 and r_3 represent the cross-lagged correlations in the model. The correlations r_1 and r_4 are static correlations which represent the relationship between grade and rating at the point in time when each set of measures was obtained.

Bohrnstedt (1969) examined the Campbell and Stanley cross-lagged panel correlation as a method of assessing causality and concluded that there are inadequacies in the technique. Bohrnstedt (1969) argues that the best predictor of Y (time 2) is Y (time 1), and correlations between X (time 1) and Y (time 2) which fail to take into account the effect of Y (time 1) will be inaccurate if not misleading. He suggests the use of gain scores in the model rather than straight time 2 measures. Gain scores are defined as the time 2 measure of a variable minus the time 1 measure of that variable. If gain scores were employed in Figure 1, the cross-lagged correlations would be computed between the time 1 measure of expected grade and the difference between the time 1 and time 2 measures of instructor rating and, between the time 1 measure of instructor rating and the difference between the time 1 and time 2 measures of expected grade. This procedure corrects for the effect of undesired time 1 variances in the cross-lagged analysis. In fact, Bohrnstedt (1969) concludes that it "overcorrects" for time 1 effects such that gain scores are negatively correlated with the time 1 measures from which they were derived.

Heise (1970) began with the work of Bohrnstedt (1969) and developed a method of analysis which more efficiently removes the effect of time 1 measures from the cross lagged correlation. He suggests path analysis as an alternative to Pearson Product Moment (PPM) correlation. The path coefficients are estimated through a series of multiple regression analyses. The standardized partial regression coefficients which result provide the values for the cross-lagged relationships (Heise 1970). Pelz and Andrews (1964) came to similar conclusions about the use of raw cross-lagged correlations and suggested the use of partial correlations rather than partial regression weights in order to estimate the values of each diagonal. When data is in standard score form, these methods produce the same results.

Another shortcoming of the cross-lagged technique is addressed by Lawler and Suttle (1972). They suggest that although the cross-lagged panel correlation has advantages over static correlation techniques, a causal relationship between two variables cannot be confirmed on the basis of this analysis alone. A significant relationship obtained using this design may be due to the influence of a third variable. In order to determine whether this has occurred, the changes in the measures of X and Y should be computed and correlated. A significant relationship between ΔX and ΔY is unlikely to be the result of the influence of a third variable. "In order for a third variable to produce a correlation between ΔX and ΔY it would not only have to change itself, but these changes would have to affect both X and Y in the same way at the same time" (Lawler and Suttle, 1972, p. 270). In this model, ΔX corresponds to the gain score defined by Bohrnstedt as the difference between time 1 and time 2 measures of variable X. Similarly, ΔY refers to the difference between time 1 and time 2 measures of variable Y. The correlation between ΔX and ΔY is referred to as a dynamic correlation, the purpose of which is to rule out the influence of extraneous variables on the obtained relationships.

Problem

Several studies have identified a relationship between students' grades (known or expected) and their ratings of instructors (Kooker, 1968; Treffinger and Feldhusen, 1970; Schuh and Crivelli, 1973; etc.). Moreover, when such a relationship is reported, it is often discussed not as a correlation, but rather as a causal truth. The purpose of this paper is to retest for the presence of a relationship between students' expected grades and their evaluations of instructors and, by employing static, dynamic, and cross-lagged panel correlations, to define more clearly the causal direction of the relationship, if it is shown to exist.

Hypotheses

A pilot study (Willoughby and Mendel, 1974) led to the formulation of the following hypotheses:

H_1. Individual students' expected grades and individual students' ratings of an instructor measured at the same point in time will be positively correlated.

H_2. The static correlation between grades and ratings at time 2 will be significantly larger than the static correlation between grades and ratings at time 1.

H_3. Students' expected grades at time 1 will be significantly correlated with instructor ratings at time 2.

H_4. The correlation between instructor rating at time 1 and student's expected grade time 2 will be significantly less than the correlation between student's expected grade at time 1 and instructor ratings at time 2.

H_5. Changes in expected grade will be positively correlated with changes in instructor rating across time.

CHAPTER II

METHOD

Sample

Fifty classes were selected during the spring semester of 1975 at Western Kentucky University. The classes were chosen as representative of classes offered by every department within the University. Sample courses ranged from freshman introductory courses to senior-graduate courses. The size of the sample classes varied from seven to ninety-six, with a median class size of thirty-one. All classes were regularly scheduled semester classes with official enrollments greater than five. Thirteen of the sample classes were taught by females; thirty-seven by males.

Questionnaire

Two course evaluation forms were employed in the study. The first, the Student Instructional Report developed by ETS, is a standardized questionnaire containing a combination of Likert and multiple-choice items. This instrument was of interest primarily because of its demonstrated reliability and validity, and the availability of national norms (Centra, 1972). The other instrument included in this study was developed by a group of Western Kentucky University faculty members. This evaluation form consists of Likert items, many of which appear to tap facets of teaching behavior not sampled by the ETS form. The resulting composite instrument had the appearance of a single 62-item questionnaire (see Appendix A). A standardized IBM answer sheet was used in all phases of data collection (see Appendix B).

Procedure

Data were collected at two points in the semester. During the pretest, conducted during the fourth week of the semester, only the fifty selected classes were measured. The post-test for these classes was conducted during the fourteenth week of an 18-week semester, concurrent with a university-wide student evaluation. Students in the fifty pretest classes were asked to identify their answer sheets by coding in their Social Security numbers. In these classes, the instructions explained that identifying numbers were necessary in order to facilitate the matching of pretest with post-test data. The instructions also assured students that their ratings would not influence their course grades, and that their instructor would never see individual evaluations. Except for requesting identification from pre- and post-test participants, the procedure for administering the evaluations was identical for classes participating in the pretest, post-test, and university-wide evaluation. In all instances, a packet containing evaluation forms, answer sheets, and instructions to the class was delivered to each participating instructor. The instructors were asked to give the unopened packet to a student in the class who could serve as monitor, and then to leave the classroom while the students completed the questionnaire. The student monitor was to read the instruction sheet aloud and distribute questionnaires and answer sheets. When the class had completed the evaluation forms, the student monitor was to collect the questionnaires and answer sheets, seal them inside the packet envelope, and deliver the sealed envelope to the secretary of the department offering the course.

Analysis

All items were transformed to Z-scores before the formal analysis was begun. An examination of item 51 (Appendix A) will help illustrate the reasoning behind the rescaling.

Item 51 is a multiple-choice item designed to measure student's expected grades. The first four response alternatives correspond with letter grades; the sixth response alternative does not and neither do the seventh and eighth. It was assumed that the fifth response alternative, "fail," corresponds with the letter grade "F." In order to facilitate the use of the response alternatives which correspond with letter grades while minimizing the impact of responses made to those alternatives which do not, all responses were transformed to _Z_ scores and response alternatives 6, 7, and 8, were defined as missing values and assigned a value of zero.

In addition to providing a conservative treatment of missing data, the _Z_ score transformation is necessary before responses made to items which have different numbers of response alternatives can be combined into an unweighted composite score. If the _Z_ score transformation was not used, items would not contribute equally to the composite score, but rather each response would be weighted by its respective item standard deviation.

When the data had been standardized, the causal relationship between expected grades and instructor rating was addressed via cross-lagged panel correlation within classes.

The cross-lagged model was applied to two pairs of variables. In the first variable pair, "instructor rating" was defined as the _Z_ score transformation of the response each student made to a single item which asks for an overall rating of the instructor (see Appendix A, item 62). The estimated reliability for this item is .85 for 15 students (Centra, 1972). "Expected grade" was similarly defined as the _Z_ score transformation of the response each student made to a single item which asks what grade the student expects to make in a course (see Appendix A, item 51). The first variable pair was called "expected grade" and "single item rating."

In the second pair of variables, expected grade was defined in the same way as it was in the first pair of variables. In addition to this variable, the second pair of variables included a subscale score rating of instructor efficiency. This score was computed as a linear composite of the _Z_ score transformations of responses each student made to items identified by the developers of the instrument as a factor representing "instructor efficiency" (see starred items, Appendix A). This factor was identified along with five other factors on the basis of a principal component factor analysis with an oblique solution, of data collected in universities across the country (Centra, 1972). The national factor structure was employed in this study in order to facilitate the comparison of these results with results obtained using the same instrument in other settings and thus maximize the generalizibility of the findings. The "instructor effectiveness" factor was selected because visual inspection of its component items suggests that it more specifically measures instructor performance than the five other factors which deal with the appropriateness of assigned readings, frequency of tests, course difficulty, lectures, and teacher-student interaction (Centra, 1972). As shown in Appendix A, responses made to the starred items were combined to form the subscale rating of instructor effectiveness.

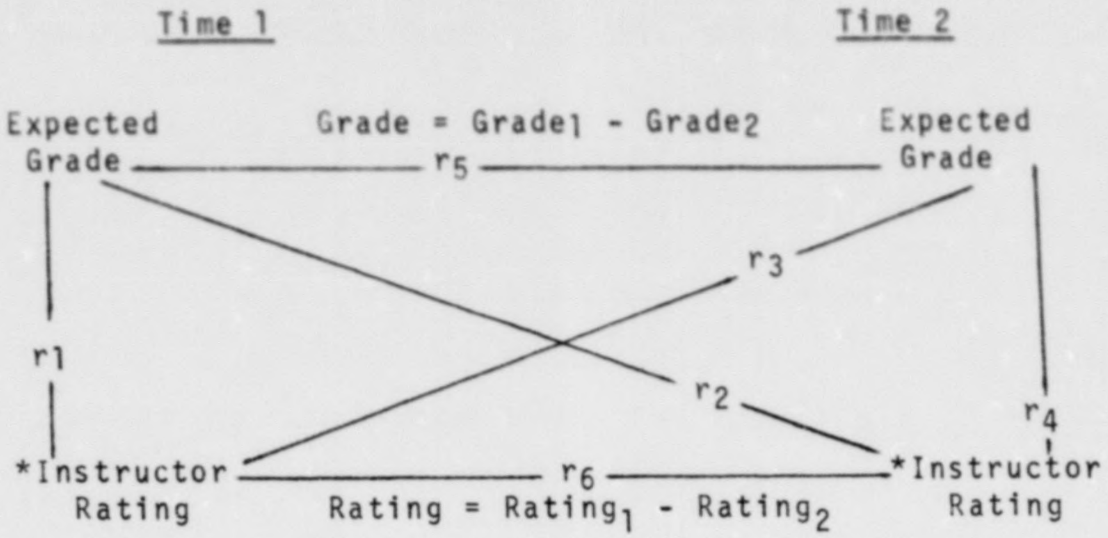
Each pair of variables (expected grade with single item rating, and expected grade with subscale rating) was analyzed using the cross-lagged technique. The pretest provided time 1 measures of each variable and the post-test provided time 2 measures. In the discussion of the cross-lagged analysis which follows, a distinction is not made between the variable pairs. The same procedures were followed when the variables were expected grade and single item rating, that were followed when the variables were expected grade and subscale rating. The complexity of the last sentence gives the reader an indication of the confusion which would result if the variable pair distinction were made throughout the discussion of cross-lagged correlations. In order to simplify the discussion, both "single item rating" and "subscale rating" are characterized as "instructor rating" in the following section. The reader should keep in mind, however, that the cross-lagged method was actually employed independently on the two variable pairs which have been identified; and results will be reported separately for the analysis of each variable pair.

The cross-lagged panel correlational model consists of three types of correlations (see Figure 2). The purpose of each and how each is computed is outlined in the sections which follow.

Static correlation

The purpose of static correlation is to provide an estimate of the relationship between two variables which are measured at the same point in time. As shown in Figure 2, the static correlations are r_1 and r_4.

The time 1 relationship between expected grade and instructor rating is represented by r_1; and the time 2



Static Correlations (PPM r)

r_1, r_4

Cross-lagged Correlations

r(2. Instructor Rating [time 1]), r_2

r(3. Expected Grade [time 1]), r_3

Dynamic Correlation (PPM r)

$r\Delta$

*Single-Item Rating or Subscale Score

Fig. 2. Applied cross-lagged correlation model

relationship between expected grade and instructor rating is shown as r_4. In the present study, the static correlations between measures of grade and rating were computed separately for students in each class. This resulted in one r_1

correlation, and one r_4 correlation for each class included in the sample. An average of these correlations was obtained through a procedure outlined by Downie and Heath (1970). The r_1 correlations for each class were converted to Fisher Zs. These Zs were weighted by N-3 (where N = the number of students in a class), and the weighted Zs were summed. An average was computed by dividing this sum by the sum of N-3 for all classes. This average Z was then converted to an r which represents the average correlation between grades and ratings at time 1. The same procedure was followed to obtain the average correlation between grades and ratings at time 2, represented by r_4.

Cross-lagged correlation

The purpose of cross-lagged correlation is to assess the direction of causality between two variables, based on repeated measures. In Figure 2, the cross-lagged correlations are represented by r_2 and r_3. The correlation between expected grade time 1 and instructor rating time 2 is represented by r_2. The correlation between instructor rating time 1 and expected grade time 2 is r_3. Some writers have suggested that cross-lagged correlations should be computed as standardized partial regression weights (Heise, 1970). When the data are standardized, this can be done by computing a partial correlation between X time 1 and Y time 2, parceling out the effects of Y time 1; and a partial correlation between Y time 1 and X time 2 holding X time 1 constant (Pelz and Andrews, 1964). Other writers have suggested that partial correlations are not necessary when the cross-lagged model is used in conjunction with dynamic correlation (Lawler and Suttle, 1972). Because there are a number of supporters of both techniques, both Pearson Product Moment correlations (PPM) and partial correlations were computed to obtain two versions of each cross-lagged correlation: PPM r and partial r holding time 1 measures constant for the time 2 variable. The reported values of r_2 and r_3 represent the average of these correlations as they were computed in each sample class (Downie and Heath, 1970).

Dynamic correlation

When cross-lagged correlations are obtained which support a causal relationship between two variables under consideration, dynamic correlation can be used to rule out the influence of a third variable not formally included in the model (Lawler and Suttle, 1972). The symbol used to represent dynamic correlation in Figure 2 is $r\Delta$. This statistic is obtained by correlating the change in expected grade from time 1 to time 2 (Δ grade) with the change in instructor rating from time 1 to time 2 (Δ rating). When the PPM correlation between Δ grade and Δ rating is significant, Lawler and Suttle (1972) suggest the influence of a variable outside the system can be ruled out with regard to the cross-lagged relationships which have been obtained. The reported values of $r\Delta$ represent the average of $r\Delta$ as it was computed in each sample class (Downie and Heath, 1970).

CHAPTER III

RESULTS

Of the fifty classes selected for pretest measurement, forty-seven classes returned the questionnaires as requested, while three classes decided not to participate. One class had inadvertently been included in the pretest sample which was scheduled to meet only during the first half of the semester. Because no post-test data could be collected for this class, the total number of classes with usable pretest data was reduced to forty-six. Students in ten of these classes apparently decided against using their Social Security numbers to identify their answer sheets, and consequently, pre- and post-data from these classes could not be matched. In an additional seven classes, pre- and post-data could not be matched due to errors in class ID codes. Of the thirty-one classes for which both pre- and postdata were available, twenty classes had more than ten students whose data were complete. The results reported in this section are based on data from these twenty classes.

In order to determine the representativeness of the sample, frequency distributions for all items were compared with frequency distributions obtained from the university as a whole. Based on this comparison, the sample was judged representative of the university population (see Table 2).

As shown in Table 3, the results support the hypothesis of a causal relationship between expected grades and single item instructor ratings. There is no evidence of a similar relationship between expected grades and subscale instructor ratings. A more detailed examination of the obtained correlations for each variable pair is presented in the following sections.

Expected Grade and Single-Item Ratings

The static correlations between expected grade and single item instructor rating are significant both at time 1 and time 2. The time 1 relationship is .15 ($p < .025$, df = 288) and the time 2 relationship is .21 ($p < .005$, df = 288). On the basis of these correlations, Hypothesis 1, that grades and ratings measured at the same point in time will be significantly related, was confirmed for this variable pair.

A *t* test of the difference between r_1 (.15) and r_4 (.21) was significant ($p < .05$, df = 288), supplying evidence

TABLE 2

FREQUENCY DISTRIBUTIONS

Single Items and Items Composing Subscale Scores										
Single Items	Sample					University				
	<u>A</u>	<u>B</u>	<u>C</u>	<u>D</u>	<u>F</u>	<u>A</u>	<u>B</u>	<u>C</u>	<u>D</u>	<u>F</u>
51. Expected Grade	38	40	16	1	-	36	39	16	1	-
	Lowest 10%	Lowest 30%	Average	Highest 30%	Highest 10%	Lowest 10%	Lowest 30%	Average	Highest 30%	Highest 10%
62. Overall Instructor Rating	4	8	25	36	26	2	9	28	32	25
		Strongly Agree	Agree	Disagree	Strongly Disagree		Strongly Agree	Agree	Disagree	Strongly Disagree
<u>Factor Component Items</u>										
1. Objectives made clear.		31	57	9	2		31	56	8	2
2. Agreement between objectives and teaching.		31	57	8	2		27	57	10	2
3. Instructor used class time well.		35	49	10	3		34	51	10	3
12. Instructor was well prepared for class.		42	50	6	1		41	47	7	2
14. Instructor summarized or emphasized major points.		34	52	11	1		32	53	9	-
20. Instructor accomplished objectives for the course.		31	60	7	1		29	57	8	-

TABLE 3

RESULTS OF CROSS-LAGGED ANALYSIS

Variable Pair	Average Correlation								
	r_1	r_2	r_2 partial	r_3	r_3 partial	r_4	$r\Delta$	r_5	r_6
Expected Grade/ Single-Item Instructor Rating	$.15^{*}$	$.30^{+}$	$.28^{+}$	$.22^{+}$	$.195^{+}$	$.21^{+}$	$.20^{+}$	$.46^{+}$	$.39^{+}$
Expected Grade/ Subscale Score Instructor Rating	$.165^{*}$	$.315^{+}$	$.285^{+}$	$.31^{+}$	$.270^{+}$	$.29^{+}$	$.12^{o}$	$.46^{+}$	$.64^{+}$

* p .025, df = 288

+ p .005, df = 288

o ns

for the confirmation of Hypothesis 2. The relationship between grades and ratings is stronger at time 2 than it is at time 1.

The direction of causality in the model was determined by examination or r_2 and r_3 . Since r_2 (.30) is significant, Hypothesis 3 that expected grades at time 1 will significantly correlate with ratings at time 2 was confirmed with regard to PPM correlations between measures of this variable pair. Because r_2 (.30) is significantly larger than r_3 (.22) ($t = 1.95$, $p < .025$, $df = 288$), Hypothesis 4 that the correlation between $grade_1$ and $rating_2$ will be significantly greater than the correlation between $rating_1$ and $grade_2$ was also confirmed based on PPM correlations for the first variable pair.

The partial correlations between (a) expected grade time 1 and single-item rating time 2 (holding single-item rating time 1 constant) and (b) single-item rating time 1 and expected grade time 2 (holding expected grade time 1 constant) are represented by r_2 partial and r_3 partial, respectively. Although r_2 partial (.28) is slightly less than r_2 (.30), and r_3 partial (.195) is less than r_3 (.22), these differences are not statistically significant. An examination of the difference between r_2 partial and r_3 partial results in a significant t of 2.341 ($p < .025$, df = 288) and the conclusions drawn on the basis of the partial correlations support those resulting from PPM analysis.

In order to rule out the influence of a third variable, $r\Delta$ was examined. Because $r\Delta$ (.20) is significant ($p < .005$, df = 288) it is highly unlikely that a third variable influenced the causal relationship which was obtained; consequently, Hypothesis 5, changes in expected grade will be positively correlated with changes in instructor rating, was confirmed.

Expected Grade and Subscale Ratings

Both of the static correlations (r_1 and r_4) between expected grade and subscale instructor rating are significant. The time 1 static correlation is .165 ($p < .025$, df = 288) and the time 2 static correlation is .29 ($p < .005$, df = 288). Based on these correlations, Hypothesis 1 was confirmed for the variable pair expected grade and subscale instructor rating. A t test of the difference between r_1 and r_4 was also significant ($t = 3.048$, $p < .005$, df = 288); therefore, Hypothesis 2 was also confirmed. The cross-lagged PPM correlations r_2 and r_3 are not significantly different for this variable pair. Individually, r_2 (.315) is significant ($p < .005$, df =

288), as is r_3 (.31). An examination of r_2 partial (.285) with respect to r_3 partial (.27) yields a similar relationship. Both correlations are significantly greater than zero, but there is no practical or statistical difference between the two. Based on these data (both r_2 and r_3, and r_2 partial and r_3 partial), Hypotheses 3 and 4 were rejected for this variable pair.

Because Hypotheses 3 and 4 were not confirmed, an examination of the dynamic correlation for this variable pair is not necessary since Hypothesis 5 addresses a moot question with regard to expected grades and subscale instructor rating.

CHAPTER IV

DISCUSSION

Several observations can be made about these data. Beginning with the results of the static correlations, there appears to be an increase in the relationship between grade expectations and instructor ratings from time 1 to time 2. This finding supports the suggestion made by Schuh and Crivelli (1973) that instructor ratings made early in a semester should show less of a relationship to the raters' expected grade than instructor ratings made late in the semester, when grade expectations are more firmly established in the minds of the raters. This trend in static correlation is supported by data from both variable pairs. In support of comments made by Treffinger and Feldhusen (1970), it appears that students in this study did not begin their classes devoid of grade expectations; or if they did, then by the fourth week of the semester (when pretest evaluations were made) they had developed a system of expectations which was salient enough to produce a significant static correlation with both a global rating of instructor efficiency and a subscale score measure of the same. Although the correlations were significantly lower at the time of the pretest, significant static correlations were nonetheless obtained from the pretest as well as the post-test data. The time 1 static correlations between expected grade and single-item rating (.15) and between expected grade and subscale rating (.165) account for less than 3 percent of the variance in instructor rating. The time 2 static correlation between expected grade and single-item rating (.21) accounts for less than 5 percent of the rating variance, and the time 2 static correlation between expected grade and subscale rating (.29) accounts for only about 9 percent of the rating variance.

Although the largest obtained static correlation accounts for less than 10 percent of the variance in instructor rating, Bassin (1974) has shown that even a small grade-rating relationship may result in significant changes in the percentile standing of instructors resulting from ratings assigned by students when instructor characteristics are statistically controlled. Because the present study employed only two measures, the relationship between expected grades and instructor rating at any point during the semester other than at the times of measurement, can only be estimated. If the relationship is linear, such that as the semester progresses, grade-rating correlations increase, then student evaluations should be administered as early in the semester as possible after students are given sufficient opportunity for observation. If the relationship is not linear, it may be possible to identify points during a semester when the relationship between ratings and expected grades accounts for the least amount of rating variance. In this case, ratings should be made at the point where the lowest relationship is obtained, provided that the students have had enough class time in which to observe their instructors, and preliminary analysis did not indicate a strong causal relationship between ratings made at the chosen time and grade expectations earlier in the semester. However, further investigation is necessary before trends in the static correlation between grades and ratings across a semester can be determined.

The static correlations discussed to this point support findings from earlier studies which were designed to measure the relationship between grades and ratings at a single point in time (Bassin, 1974; Treffinger and Feldhusen, 1970; Schuh and Crivelli, 1974; Caffrey, 1970, etc.). Additionally, the present study goes beyond the static correlation between grades and ratings to address the issue of causality using the cross-lagged panel correlation technique. Because there is some question as to the best statistic to employ within the cross-lagged model, both Pearson Product Moment correlations and partial correlations were computed for both variable pairs.

The cross-lagged PPM correlations were consistently higher than the cross-lagged correlations computed as partial correlations with extraneous time 1 measures statistically removed. However, in no cases were the PPM and partial correlations significantly different. There are two possible explanations for this finding. First, although significant correlations were obtained between measures of each variable, these correlations were quite low. Additionally, the averaging procedure followed in order to compute all reported values, effectively cancels extreme correlations so that the correlation coefficients which result for both partial and PPM correlations are modified representations of the correlation distributions obtained across all classes. Because the sample was chosen to be representative of the university as a whole, there was a great deal of variance between classes taught in different colleges and departments. For example, the obtained raw correlations, r_2, between single-item rating and expected grade ranged from .11 to .57. Between the same variables, the raw partial correlations, r_2 partial, ranged from .01 to .46 (see Table 4).

The data support a causal relationship between expected grades and single-item instructor ratings which, although moderate, is nonetheless statistically significant. In the case of these data, expected grades at time 1 are clearly causal contributors (not major determinants) to instructor ratings at time 2. When the definition of instructor rating is a score on a subscale of the instrument, this relationship is not supported. This is not to say that grade expectations at time 1 are not related to subscale ratings at time 2; they are. However, this relationship is not significantly stronger than the relationship between subscale rating (time 1) and expected grade (time 2). One interpretation of these results suggests that as the measures of each variable become more well-defined (e.g., factor scores as opposed to single-item ratings) there is more opportunity to assess which components of teacher performance are influenced by which rating determinants. Just as the "total score" on an evaluation form reflects components of

TABLE 4

OBTAINED PEARSON PRODUCT MOMENT CORRELATIONS

	Factor Score							Single Item						
N-3	1	2	3	4	5	6	Δ	1	2	3	4	5	6	Δ
44	.46	.10	.08	.09	.30	.42	.10	.46	.23	.16	.50	.30	.04	.11
11	.71	.24	.57	.73	.89	.50	.04	.71	.30	.09	.29	.32	.04	.07
7	.58	.24	.22	.04	.07	.77	.395	.58	.13	.41	.81	.25	.13	.10
8	.25	.09	.59	.33	.11	.92	.54	.25	.33	.29	.13	.11	.02	.19
12	.65	.17	.63	.17	.48	.51	.09	.65	.27	.04	.14	.265	.09	.11
8	.27	.04	.33	.28	.22	.96	.395	.27	.305	.20	.11	.10	.12	.04
25	.48	.08	.33	.13	.30	.40	.13	.48	.38	.13	.29	.27	.03	.02
11	.03	.03	.31	.35	.18	.72	.33	.03	.57	.45	.50	.30	.32	.08
8	.33	.24	.59	.11	.09	.80	.10	.33	.11	.34	.47	.29	.44	.31
19	.59	.08	.43	.38	.03	.84	.16	.59	.27	.19	.09	.35	.13	.22
11	.73	.12	.085	.31	.35	.89	.07	.73	.57	.32	.09	.13	.145	.05
19	.62	.02	.38	.59	.21	.71	.31	.62	.38	.10	.19	.15	.10	.20
13	.00	.25	.34	.47	.08	.33	.06	.00	.20	.52	.16	.12	.39	.20
8	.35	.37	.24	.04	.25	.83	.14	.35	.13	.10	.93	.33	.37	.20
21	.48	.08	.08	.35	.26	.43	.19	.48	.17	.01	.27	.17	.22	.09
19	.25	.43	.38	.13	.475	.445	.30	.25	.26	.32	.55	.07	.20	.06
12	.63	.11	.06	.13	.10	.89	.26	.63	.15	.10	.52	.22	.07	.10
19	.71	.21	.02	.08	.27	.59	.38	.71	.27	.17	.12	.01	.16	.08
7	.11	.07	.05	.56	.17	.39	.17	.11	.33	.09	.40	.11	.20	.23
7	.04	.55	.40	.25	.33	.70	.00	.04	.33	.15	.27	.35	.19	.20

the class not related to instructor performance, so might a single-item rating of overall instructor performance reflect components of teaching behavior other than those included in the relatively discrete measure represented by a factor score. In the present study, that part of the single-item rating which is not shared by the subscale rating seems to be causally influenced by students' expected grades.

The dynamic correlation results have been reviewed earlier. The conclusions drawn at that point stated that between single-item ratings and expected grades, the causal relationship obtained can be considered significant. Changes in global ratings correspond with changes in grade expectations to such an extent that outside variable influences can be discounted. Between subscale ratings and expected grades, however, this correlation is nonsignificant. Since the cross-lagged model failed to indicate causality between subscale ratings and expected grades, this failure to reach significance was not unexpected.

From these data it can be determined that the subscale rating of instructor effectiveness and the single-item rating measure different aspects of instructor performance. The correlation between subscale ratings and single-item ratings, averaged across all twenty classes, was of moderate strength and statistically significant ($r = .24$, $p < .05$, $df = 288$); however, the correlation indicates that less than 5 percent of the total rating variance is shared by the two measures of instructor performance.

Because the single-item rating represents a global assessment of instructor performance, it may reflect many varying characteristics of teaching behavior. In order to determine how a given instructor ranks among his peers, student raters might consider many aspects of the instructor's professional performance including the instructor's accessibility, subject matter competence, ability to gauge students understanding, fairness, etc. The subscale rating, on the other hand, reflects only one aspect of teacher behavior, Instructional Efficiency, or the clarity with which an instructor organizes and presents material to the class. On the basis of this study, it can be shown that students' expected grades causally influence their overall evaluation of an instructor but not their rating of an instructor on a somewhat more objective, and certainly more discrete subscale. Based on these findings, one way to minimize the causal bias associated with students' grade expectations is to objectify the evaluation as much as possible. If questions _are_ included which draw subjective student responses, then these ratings should be accompanied by an explanation of the possible influence of what Schuh and Crivelli (1974) called "animadversion error."

Although the preceding discussion has been based on correlations which are statistically significant, the practical significance of these correlations is open to question, and should be discussed. Due to the large sample size, correlations as low as .16 are judged significantly different than zero. In the case of the first variable pair, the causal relationship confirmed between grades and ratings was based on such a correlation. Although .28 (r_2) is significantly greater than .195 (r_3), the larger correlation accounts for only about 7 percent of the variance in instructor ratings. In fact, the largest obtained correlation (.315, r_2 for the second variable pair) accounts for less than 10 percent of the total variance in instructor ratings.

Bassin (1974) has shown that even a small relationship between grades and ratings can result in significant rating changes when grades are artifically manipulated. However, it is evident that students' expected grades are _not_ the major determinants of instructor rating. In individual classes, grade expectations account for greater or lesser amounts of instructor rating variance, based on the characteristics of each class. The range of correlations obtained in this study is evidence for this point (see Table 4).

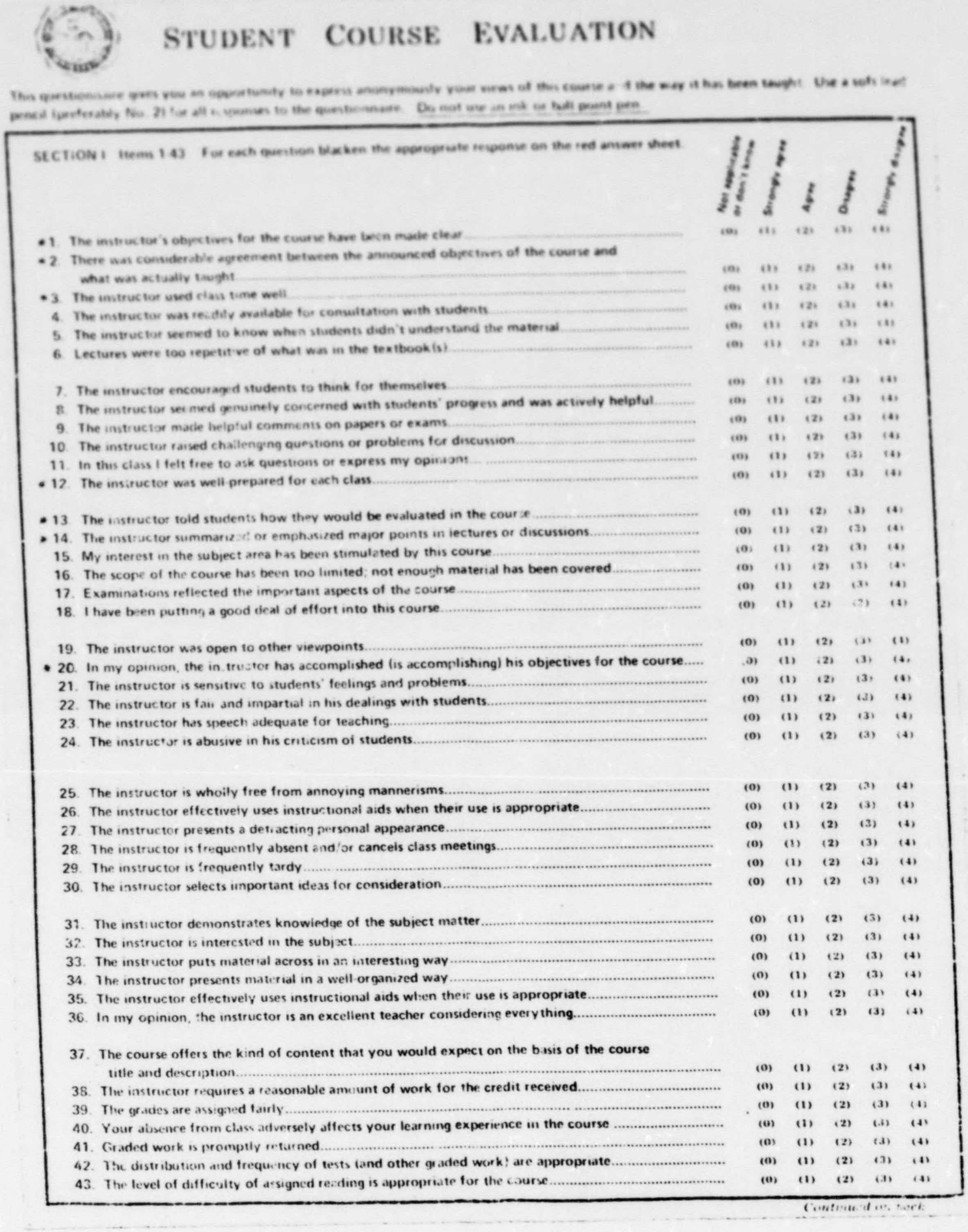
The results of this study represent an average across widely differing classes within the university and, consequently, they represent the individual results from some classes better than others. Future studies in the area might group classes on the basis of common characteristics in order to reduce the diversity obtained when classes are chosen to represent an entire university. Additionally, other subscale scores should be entered into the cross-lagged model along with other potential determinants of each of these ratings.

After studies have been done to isolate more components of student evaluations, these evaluations will be more useful in that they will be interpretable within the method's limitations. Moderator variables may be identified which isolate groups of raters who consistently are able to provide ratings that are relatively free from bias. When this state is reached, instructors will be assured that information obtained through the administration of such evaluations is valid, and can be useful to them, rather than a source of suspicion and/or disdain.

Ideally, even the student-raters will be provided with summary information regarding the outcome of their endeavors. These changes will only be realized, however, when the student evaluation itself is better understood.

APPENDIX A

QUESTIONNAIRE



STUDENT COURSE EVALUATION

This questionnaire gives you an opportunity to express anonymously your views of this course and the way it has been taught. Use a soft lead pencil (preferably No. 2) for all responses to the questionnaire. Do not use an ink or ball point pen.

SECTION I Items 1-43 For each question blacken the appropriate response on the red answer sheet.

	Not applicable or don't know	Strongly agree	Agree	Disagree	Strongly disagree
* 1. The instructor's objectives for the course have been made clear	(0)	(1)	(2)	(3)	(4)
* 2. There was considerable agreement between the announced objectives of the course and what was actually taught	(0)	(1)	(2)	(3)	(4)
* 3. The instructor used class time well	(0)	(1)	(2)	(3)	(4)
4. The instructor was readily available for consultation with students	(0)	(1)	(2)	(3)	(4)
5. The instructor seemed to know when students didn't understand the material	(0)	(1)	(2)	(3)	(4)
6. Lectures were too repetitive of what was in the textbook(s)	(0)	(1)	(2)	(3)	(4)
7. The instructor encouraged students to think for themselves	(0)	(1)	(2)	(3)	(4)
8. The instructor seemed genuinely concerned with students' progress and was actively helpful	(0)	(1)	(2)	(3)	(4)
9. The instructor made helpful comments on papers or exams	(0)	(1)	(2)	(3)	(4)
10. The instructor raised challenging questions or problems for discussion	(0)	(1)	(2)	(3)	(4)
11. In this class I felt free to ask questions or express my opinions	(0)	(1)	(2)	(3)	(4)
* 12. The instructor was well prepared for each class	(0)	(1)	(2)	(3)	(4)
* 13. The instructor told students how they would be evaluated in the course	(0)	(1)	(2)	(3)	(4)
* 14. The instructor summarized or emphasized major points in lectures or discussions	(0)	(1)	(2)	(3)	(4)
15. My interest in the subject area has been stimulated by this course	(0)	(1)	(2)	(3)	(4)
16. The scope of the course has been too limited; not enough material has been covered	(0)	(1)	(2)	(3)	(4)
17. Examinations reflected the important aspects of the course	(0)	(1)	(2)	(3)	(4)
18. I have been putting a good deal of effort into this course	(0)	(1)	(2)	(3)	(4)
19. The instructor was open to other viewpoints	(0)	(1)	(2)	(3)	(4)
* 20. In my opinion, the instructor has accomplished (is accomplishing) his objectives for the course	(0)	(1)	(2)	(3)	(4)
21. The instructor is sensitive to students' feelings and problems	(0)	(1)	(2)	(3)	(4)
22. The instructor is fair and impartial in his dealings with students	(0)	(1)	(2)	(3)	(4)
23. The instructor has speech adequate for teaching	(0)	(1)	(2)	(3)	(4)
24. The instructor is abusive in his criticism of students	(0)	(1)	(2)	(3)	(4)
25. The instructor is wholly free from annoying mannerisms	(0)	(1)	(2)	(3)	(4)
26. The instructor effectively uses instructional aids when their use is appropriate	(0)	(1)	(2)	(3)	(4)
27. The instructor presents a detracting personal appearance	(0)	(1)	(2)	(3)	(4)
28. The instructor is frequently absent and/or cancels class meetings	(0)	(1)	(2)	(3)	(4)
29. The instructor is frequently tardy	(0)	(1)	(2)	(3)	(4)
30. The instructor selects important ideas for consideration	(0)	(1)	(2)	(3)	(4)
31. The instructor demonstrates knowledge of the subject matter	(0)	(1)	(2)	(3)	(4)
32. The instructor is interested in the subject	(0)	(1)	(2)	(3)	(4)
33. The instructor puts material across in an interesting way	(0)	(1)	(2)	(3)	(4)
34. The instructor presents material in a well-organized way	(0)	(1)	(2)	(3)	(4)
35. The instructor effectively uses instructional aids when their use is appropriate	(0)	(1)	(2)	(3)	(4)
36. In my opinion, the instructor is an excellent teacher considering everything	(0)	(1)	(2)	(3)	(4)
37. The course offers the kind of content that you would expect on the basis of the course title and description	(0)	(1)	(2)	(3)	(4)
38. The instructor requires a reasonable amount of work for the credit received	(0)	(1)	(2)	(3)	(4)
39. The grades are assigned fairly	(0)	(1)	(2)	(3)	(4)
40. Your absence from class adversely affects your learning experience in the course	(0)	(1)	(2)	(3)	(4)
41. Graded work is promptly returned	(0)	(1)	(2)	(3)	(4)
42. The distribution and frequency of tests (and other graded work) are appropriate	(0)	(1)	(2)	(3)	(4)
43. The level of difficulty of assigned reading is appropriate for the course	(0)	(1)	(2)	(3)	(4)

Continued on back

SECTION II Items 44-54 For each question blacken the appropriate response number on the red answer sheet.

44. For my preparation and ability, the level of difficulty of this course was:

(1) Very elementary
(2) Somewhat elementary
(4) Somewhat difficult
(5) Very difficult

45. The work load for this course in relation to other courses or equal credit was:

(1) Much lighter
(2) Lighter
(3) About the same
(4) Heavier
(5) Much Heavier

46. Was class size satisfactory for the method of conducting the class?

(1) Yes, most of the time
(2) No, class was too large
(3) No, class was too small
(4) It didn't make any difference one way or the other

47. Which one of the following best describes this course for you?

(1) Major requirement or elective within major field
(2) Minor requirement or required elective outside major field
(3) College requirement but not part of my major or minor field
(4) Elective not required in any way
(5) Other

48. Which one of the following was your most important reason for selecting this course?

(1) Friend(s) recommended it
(2) Faculty advisor's recommendation
(3) Teacher's excellent reputation
(4) Thought I could make a good grade
(5) Could use pass/no credit option
(6) It was required
(7) Subject was of interest
(8) Other

49. For me, the pace at which the instructor covered the material during the term was:

(1) Very slow
(2) Somewhat slow
(3) Just about right
(4) Somewhat fast
(5) Very fast

50. To what extent did the instructor use examples or illustrations to help clarify the material?

(1) Never
(2) Seldom
(3) Occasionally
(4) Frequently

51. What grade do you expect to receive in this course?

(1) A
(2) B
(3) C
(4) D
(5) Fail
(6) Pass
(7) No credit
(8) Other

52. What is your approximate cumulative grade-point average?

(1) 3.50 - 4.00
(2) 3.00 - 3.49
(3) 2.50 - 2.99
(4) 2.00 - 2.49
(5) 1.50 - 1.99
(6) 1.00 - 1.49
(7) Less than 1.00
(8) None yet—freshmen or transfer

53. What is your class level?

(1) Freshman
(2) Sophomore
(3) Junior
(4) Senior
(5) Graduate
(6) Other

54. Sex:

(1) Female
(2) Male

SECTION III Items 55-62 For each question blacken the appropriate response on the red answer sheet.

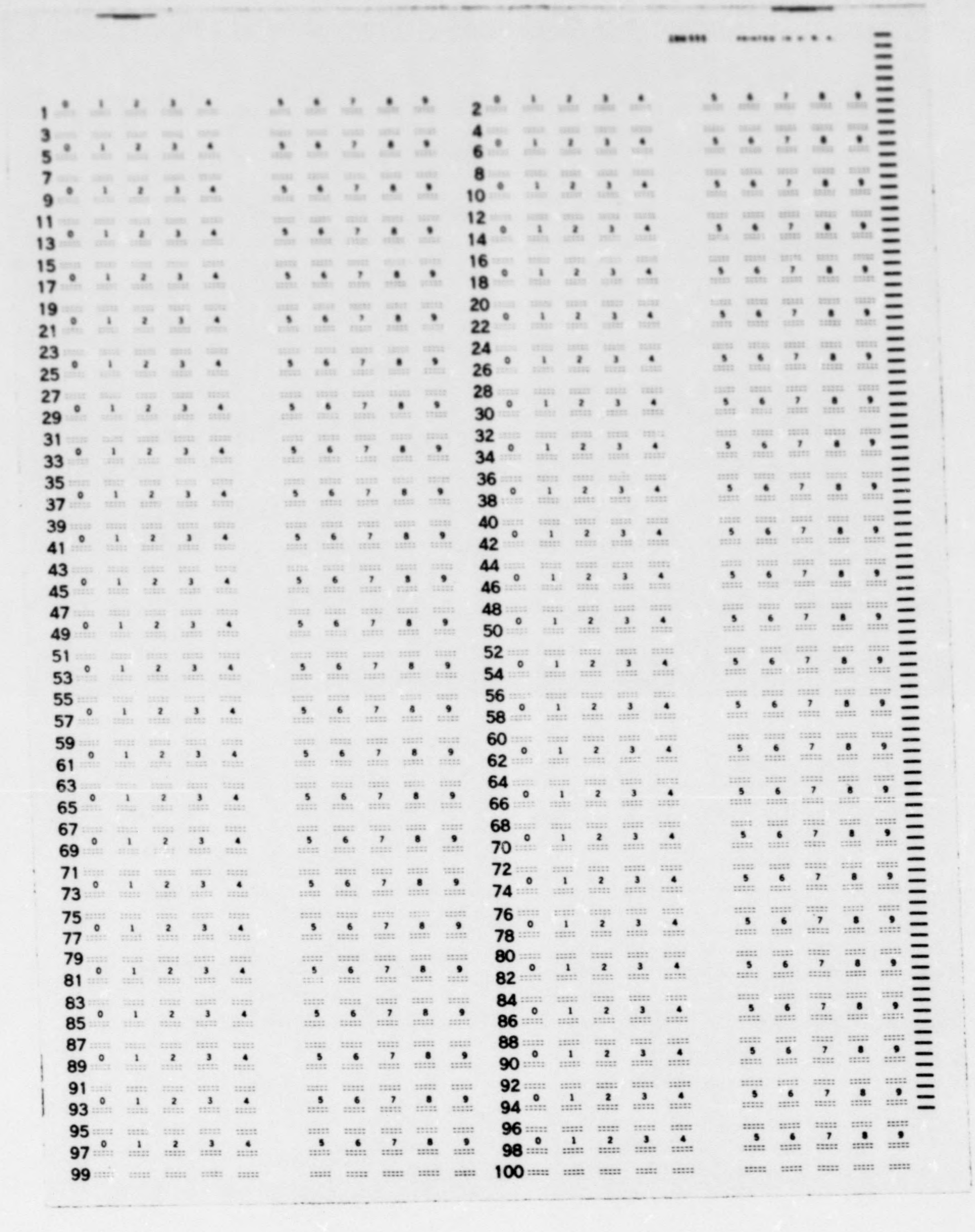
	Not applicable, don't know, or there were none	Poor	Fair	Satisfactory	Good	Excellent
55. Overall, I would rate the textbook(s)..	(0)	(1)	(2)	(3)	(4)	(5)
56. Overall, I would rate the supplementary readings..	(0)	(1)	(2)	(3)	(4)	(5)
57. Overall, I would rate the quality of the exams..	(0)	(1)	(2)	(3)	(4)	(5)
58. I would rate the general quality of the lectures..	(0)	(1)	(2)	(3)	(4)	(5)
59. I would rate the overall value of class discussions..	(0)	(1)	(2)	(3)	(4)	(5)
60. Overall, I would rate the laboratories..	(0)	(1)	(2)	(3)	(4)	(5)
61. I would rate the overall value of this course to me as..	(0)	(1)	(2)	(3)	(4)	(5)

62. Compared to other instructors you have had (secondary school and college), how effective has the instructor been in this course? (Blacken one response number.)

One of the least effective (among the lowest 10%)	Not as effective as most (in the lowest 30%)	About Average	More effective than most (among the top 30%)	One of the most effective (among the top 10%)
(1)	(2)	(3)	(4)	(5)

APPENDIX B

ANSWER SHEET



IBM 555 PRINTED IN U.S.A.

1	0 1 2 3 4 5 6 7 8 9	2	0 1 2 3 4 5 6 7 8 9
3		4	
5	0 1 2 3 4 5 6 7 8 9	6	0 1 2 3 4 5 6 7 8 9
7		8	
9	0 1 2 3 4 5 6 7 8 9	10	0 1 2 3 4 5 6 7 8 9
11		12	
13	0 1 2 3 4 5 6 7 8 9	14	0 1 2 3 4 5 6 7 8 9
15		16	
17	0 1 2 3 4 5 6 7 8 9	18	0 1 2 3 4 5 6 7 8 9
19		20	
21	0 1 2 3 4 5 6 7 8 9	22	0 1 2 3 4 5 6 7 8 9
23		24	
25	0 1 2 3 4 5 6 7 8 9	26	0 1 2 3 4 5 6 7 8 9
27		28	
29	0 1 2 3 4 5 6 7 8 9	30	0 1 2 3 4 5 6 7 8 9
31		32	
33	0 1 2 3 4 5 6 7 8 9	34	0 1 2 3 4 5 6 7 8 9
35		36	
37	0 1 2 3 4 5 6 7 8 9	38	0 1 2 3 4 5 6 7 8 9
39		40	
41	0 1 2 3 4 5 6 7 8 9	42	0 1 2 3 4 5 6 7 8 9
43		44	
45	0 1 2 3 4 5 6 7 8 9	46	0 1 2 3 4 5 6 7 8 9
47		48	
49	0 1 2 3 4 5 6 7 8 9	50	0 1 2 3 4 5 6 7 8 9
51		52	
53	0 1 2 3 4 5 6 7 8 9	54	0 1 2 3 4 5 6 7 8 9
55		56	
57	0 1 2 3 4 5 6 7 8 9	58	0 1 2 3 4 5 6 7 8 9
59		60	
61	0 1 2 3 4 5 6 7 8 9	62	0 1 2 3 4 5 6 7 8 9
63		64	
65	0 1 2 3 4 5 6 7 8 9	66	0 1 2 3 4 5 6 7 8 9
67		68	
69	0 1 2 3 4 5 6 7 8 9	70	0 1 2 3 4 5 6 7 8 9
71		72	
73	0 1 2 3 4 5 6 7 8 9	74	0 1 2 3 4 5 6 7 8 9
75		76	
77	0 1 2 3 4 5 6 7 8 9	78	0 1 2 3 4 5 6 7 8 9
79		80	
81	0 1 2 3 4 5 6 7 8 9	82	0 1 2 3 4 5 6 7 8 9
83		84	
85	0 1 2 3 4 5 6 7 8 9	86	0 1 2 3 4 5 6 7 8 9
87		88	
89	0 1 2 3 4 5 6 7 8 9	90	0 1 2 3 4 5 6 7 8 9
91		92	
93	0 1 2 3 4 5 6 7 8 9	94	0 1 2 3 4 5 6 7 8 9
95		96	
97	0 1 2 3 4 5 6 7 8 9	98	0 1 2 3 4 5 6 7 8 9
99		100	

BIBLIOGRAPHY

BIBLIOGRAPHY

Anikeeff, A. M. Factors affecting student evaluation of college faculty members. Journal of Applied Psychology, 1953, 37, 458-460.

Bassin, W. M. A note on the biases in students' evaluations of instructors. Journal of Experimental Education, 1974, 43, 16-18.

Bendig, A. W. Relationship of level of course achievement of students, instructor and course ratings in introductory psychology classes. Educational and Psychological Measurement, 1953, 13, 437-448. (a)

Bendig, A. W. Student achievement in introductory psychology and student ratings of the competence and empathy of their instructors. Journal of Psychology, 1953, 36, 427-433. (b)

Blum, M. L. An investigation of the relationship between students grades and their ratings of an instructors ability to teach. Journal of Educational Psychology, 1936, 27, 217-221.

Bohrnstedt, George W. Observations on the measurement of change. Sociological methodology. San Francisco: Jonsey-Bass, Inc., 1969.

Caffrey, B. Lack of bias in student evaluation of teachers. Proceedings of the 77th Annual Convention of the American Psychological Association, 1969, 4, 641-642.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Centra, J. A. The student instructional report no. 3. Princeton, N.J.: ETS, College and University Programs, 1972.

Centra, J. A. Comparison of teacher self-ratings with student ratings. Journal of Educational Psychology, 1973, 287-295.

Costin, F. C., Greenough, W. T., & Menges, R. J. Student ratings of college teaching: Reliability, validity and usefulness. Review of Educational Research, 1971, 41, 511-535.

Downie, N. M., & Heath, R. W. Basic statistical methods (3rd ed.). New York: Harper & Row, 1970.

Echandia, P. P. A methodological study and factor-analytic validation of forced choice performance ratings of college accounting instructors. Dissertation Abstracts International, 1964, 25, 2605-2606.

Eckert, R. E. Ways of evaluating college teaching. School and Society, 1950, 71, 65-69.

Gaverick, C. M., & Carter, H. D. Instructor ratings and expected grades. California Journal of Educational Research, 1962, 13, 218-221.

Gustad, J. W. Policies and practices in faculty evaluation. Educational Record, 1961, 42, 194-211.

Guthrie, E. R. The evaluation of teaching. Educational Record, 1949, 30, 109-115.

Guthrie, E. R. The evaluation of teaching: A progress report. Seattle: University of Washington, 1954.

Hall, V. C. Former student evaluation as a criterion of teaching success. Journal of Experimental Education, 1965, 34, 1-19.

Heilman, J. D., & Armentrout, W. D. The rating of college teachers on ten traits by their students. Journal of Educational Psychology, 1936, 27, 197-216.

Heise, D. R. Causal inference from panel data. Sociological methodology. San Francisco: Jonsey-Bass, 1970.

Kohlan, R. G. Comparison of faculty evaluations early and late in the course. Journal of Higher Education, 1973, 44, 587-595.

Kooker, E. W. The relationship of known college grades to course ratings on student-selected items. Journal of Psychology, 1968, 69, 209-215.

Lawler, E. E., & Suttle, J. L. A causal correlational test of the need hierarchy concept. Organizational Behavior and Human Performance, 1972, 7, 265-287.

Lovell, G. D., & Haner, C. F. Forced choice applied to college faculty rating. Educational and Psychological Measurement, 1955, 15, 291-304.

McKeachie, W. J. Student ratings of faculty. American Association of University Professors Bulletin, 1969, 55, 439-444.

McKeachie, W. J., & Soloman, D. Student ratings of instructors. Journal of Educational Research, 1958, 51, 319-382.

Mueller, F. J. Trends in student rating of faculty. American Association of University Professors Bulletin, 1951, 37, 319-324.

Pelz, D. C., & Andrews, F. M. Detecting causal priorities in panel study data. American Sociological Review, 1964, 29, 836-848.

Remmers, H. H. The relationship between students marks and students attitudes toward instructors. School and Society, 1928, 28, 759-760.

Remmers, H. H. To what extent do grades influence student ratings of instructors? Journal of Educational Research, 1930, 21, 314-316.

Remmers, H. H., & Drucker, A. J. Do alumni and students differ in their attitudes toward instructors? Journal of Educational Psychology, 1951, 42, 129-143.

Remmers, H. H., Martin, F. D., & Elliott, D. N. Are students ratings of instructors related to their grades? Purdue University Studies in Higher Education, 1949, 66, 17-26.

Russell, H. E. & Bendig, A. W. Investigations of the relation of student ratings of psychology instructors to their course achievement when academic aptitude is controlled. Educational and Psychological Measurement, 1954, 13, 626-635.

Schuh, A. J., & Crivelli, M. A. Animadversion error in student evaluation of faculty teaching effectiveness. Journal of Applied Psychology, 1973, 58, 259-260.

Treffinger, D. J., & Feldhusen, J. F. Predicting students' ratings of instruction. Proceedings of the 78th Annual Convention of the American Psychological Association, 1970, 5, 621-622.

Tryon, R. C. Cumulative commonality cluster analysis. Educational and Psychological Measurement, 1958, 18, 3-35.

Voeks, V. W., & French, G. M. Are student-ratings of teachers affected by grades? Journal of Higher Education, 1960, 31, 330-334.

Weaver, C. H. Instructor rating by college students. Journal of Educational Psychology, 1960, 51, 21-25.

Webb, N. B., & Nolan, C. Y. Student, supervisor and self ratings of instructor proficiency. Journal of Educational Psychology, 1955, 46, 42-46.

Webster's Third International Dictionary. Springfield, Mass.: G. & C. Merriam Company, 1961.

Willoughby, D., & Mendel, R. M. A causal analysis of animadversion error. Unpublished study. Western Kentucky University, 1974.

B12, F5