1982

An analysis (between and within classrooms) of student evaluation of instruction

Benchalak Phutinart
Iowa State University

Phutinart, Benchalak

AN ANALYSIS (BETWEEN AND WITHIN CLASSROOMS) OF STUDENT EVALUATION OF INSTRUCTION

Iowa State University PH.D. 1982

An analysis (between and within classrooms) of student evaluation of instruction

by

Benchalak Phutinart

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Department: Professional Studies in Education
Major: Education (Higher Education)

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1982

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Student evaluation of faculty teaching has been a practice in universities for more than 50 years (Kent, 1966). This can be attributed to two major factors: the factors of students to have a greater role in the learning process, and the demand by the public for greater accountability within academic institutions (Tetenbaum, 1977). As a result of these demands, universities are becoming more involved in systematic research programs designed to develop reliable and valid student rating forms of faculty teaching (Aleamoni and Spencer, 1973; Barnoski and Sockloff, 1976).

Student ratings are judged to be a valuable tool for evaluation of faculty. Students provide a "ready-made" pool of observers who provide a wealth of data pertaining to such aspects of teaching as competency, organization, attitude toward students and knowledge of subject area (Graham, 1973). Secondly, as to the subjects which instructors attempt to teach, students are also in a better position to provide feedback concerning teacher effectiveness (Guthrie, 1949; McKeachie, 1975b). Compared to other sources of evaluation such as faculty self-ratings and peer-ratings, student-ratings are found to be equally as good, if not superior (Gordon et al., 1973; Doyle and Crichton, 1978). Doyle and Crichton (1978) found that the results from these three different sources of evaluation were similar in range and distribution, although colleagues tended to give the most favorable ratings, students the least favorable. Self-ratings and student ratings were found to be very similar in patterns of strengths and weaknesses across items (Centra, 1973). When student ratings are compared to the other ratings, their ratings have some strengths,

such as being summarized economically, recordable, comparable with certain other information, and the ability to have rating scale items and data studied and refined (Doyle, 1975). Gange (1974) stated that student ratings are logically justified because students have a significant amount of relevant contact with the instructor. No one sees and hears and reads and experiences teachers' work so fully, directly and personally as the teacher's students. The literature suggests that student ratings are valid measures of teaching effectiveness when compared with student learning performance (Johnson, 1976; Grush and Costin, 1975; Frey, 1973; French-Lazovik, 1974). In addition to serving as a means of evaluating teaching effectiveness, student ratings also reflect the students' educational attitudes and educational value which provide helpful information to faculty and administrators for the assessment of existing programs as well as future planning (Johnson, 1976; Stumpf, 1979).

The importance of student ratings is rapidly gaining acceptance. According to a survey conducted by the American Council on Education (ACE) in 1961, 48% of institutions did not use student ratings at all in making decisions pertaining to their faculty (Astin and Lee, 1967). A later survey, by Bejar in 1975, revealed that the figure has decreased to a mere 13%.

Increased interest in using student ratings naturally has led to increased research. Results of much of this research have appeared in literature. However, many of them still need further clarification, and some need further verification. Many problems remain to be solved, and, indeed, to be defined (Nadeau, 1974).

The researcher has found that, in many studies, the data taken from student rating forms have not been analyzed by the appropriate type of analysis. It is known that there are three types of data from student ratings that can be used for research: the total-class data, the between-class data, and the within-class data. Therefore, there are three types of analysis that can be conducted: total-class data analysis (T-analysis), the within-class data analysis (W-analysis) and the between-class data analysis (B-analysis).

In light of the fact that the instructor is the subject whom the student rating is intended to evaluate, the B-analysis is the most appropriate analysis to use for student ratings because it uses the instructor as the unit of analysis. In contrast, the T-analysis uses the individual student as the unit of analysis, treating students from all classes as one group and disregarding courses and/or section differences. Since these kinds of data do not pertain to individual instructors, it is not relevant to the instructors rated. Consequently, the total-class data analysis is inappropriate. The W-analysis is also individual student-based in approach, but the information it provides is unambiguous as contrasted with the T-analysis. The within-class correlations are based on individual differences where classroom differences are controlled. Such correlations are conceptually the correlation within a single classroom or the average of correlations over classrooms. This is, therefore, more appropriate compared to total-class correlations which confounded the two sources of covariance of the within-class covariance and the between-class covariance (McNemar, 1962; Barnoski and Sockloff, 1976; Cronbach et al., 1972; Sockloff, 1975).

In past research, most analyses were made using either total-class data or between-class data alone. The results from such studies are found to have many inconsistencies. These inconsistencies may conform first, using the T-analysis which gives the ambiguous results since it is confounded; and second, using the B-analysis alone. Consequently, studies based on the combination of different types of data or using more than one type of analysis in a single study have emerged (e.g., Bejar and Doyle, 1974; Doyle and Whitely, 1974; Whitely and Doyle, 1976; 1978, 1979; Stumpf and Freedman, 1979).

Literature concerning combination analysis using all three types of data analysis or using either T-analysis and W-analysis or using only T-analysis and B-analysis or B-analysis and W-analysis has recently been published.

The risk involved in using the T-analysis is that scores derived from such an analysis will be attributed to the teaching behavior of that instructor, whereas, in fact, those scores may reflect the attributes of that instructor's students. Student attributes such as sex, male or female, class rank and academic ability determine, to some extent, how the students rate their instructors. This risk is also present in scores derived from the B-analysis, since different instructors have different kinds of students.

Recognizing the danger of confounding data in the use of T-analysis and the need to use more than one type of data analysis, this researcher chose to carry out analyses using the between-class data and the within-class data. Presently, there has been only one other study using this

methodology, that of Whitely and Doyle (1979). Unfortunately, due to the inadequate size of the sample taken, the results of that study are not generalizable.

There are many aspects of student ratings which previous studies have shown to have inconsistent results (see review by Costin et al., 1971). Such inconsistencies have been attributed to the use of inappropriate data analysis (T-analysis) and the use of B-analysis alone, as previously discussed, using a small group and/or only one particular course and/or only one particular student level (Brandenburge et al., 1977; Graham, 1973). In addition, inferences derived from using only introductory courses should be counted as limitations of previous research. Many previous studies have favored using introductory courses as the basis for study. For example, Russell and Bendig (1953), Holmes (1972) and Bendig (1953) used an introduction to psychology class; Frey et al. (1975) used an introduction to calculus class; Whitely and Doyle used a beginning French class in 1976, in 1978, a beginning calculus class, and in 1979, a beginning mathematics class. The use of introductory courses as samples limits the study to lower student level (freshmen and sophomores). Upper students may be enrolled in introductory courses, but their number is limited, and they are more likely to be taking these courses as electives rather than as required courses. Student level and required or elective courses are variables that affect the student ratings (see review, Aleamoni and Graham, 1974; Feldman, 1978; Aleamoni and Hexner, 1980).

The present research was intended to analyze student ratings in a manner different from previous studies by first employing the B-analysis

and the W-analysis, and second, analyzing instructors from several courses, each of which is a multiple section course taught by different instructors in a variety of disciplines. In other words, the present study will analyze the student ratings on both instructor level (through the B-analysis) and student level (through the W-analysis). At the instructor level, not only are the different instructors of a course studied, but also a variety of courses from a wide range of disciplines. The purposes of this research were to:

1) Study how the variances in student rating are related to student characteristics (such as sex, course, expected grade, GPA, etc.);
2) Study the relationship between student ratings and the different courses being taken; and
3) Provide better models and methods for analysis of student ratings of instruction.

The factors to be measured will be based on the B-analysis when instructors are the unit of analysis. The results will be reported with the consideration of the results from both the B-analysis and the W-analysis to ensure maximum accuracy of the report. This means that the study will have looked at both the instructor level and the student level.

The outline for the study will be as follows:

1) Chapter one is the Introduction;
2) Chapter two presents a review of the literature with primary emphasis on previous research which uses the analysis of the between-class data;
3) Chapter three will focus on the methodology of the study;

4) Chapter four presents the results with explanations for each objective as established;

5) Chapter five is the final chapter, which presents conclusions and discussion. Interpretation and implementation are included.

REVIEW OF LITERATURE

The present study is directed toward the analysis of student evaluations of the instructor using both between-class data and within-class data. Only one research report was found in the literature (Whitely and Doyle, 1979) which used the same type of data analysis as the present study. Most previous research used a single type of analysis, either total-class data analysis (T-analysis) or between-class data analysis (B-analysis). Since Whitely and Doyle's study (1979) is the only one that used the methodology employed for this study, an extensive review of completely relevant literature is not possible. Therefore, the review of literature in this chapter will be primarily the review of previous studies which used the B-analysis. However, the review may touch somewhat on previous studies which used the T-analysis, but only whenever it is necessary to support the study.

The review is divided into several parts. Each part is for a particular part of the study and it is connected to another part of the study.

The first part of the review will focus on the three types of analysis. This should provide the reader with:

1) Better understanding of the distinction between these three types of analysis (T-analysis, B-analysis and W-analysis);
2) Why the present study abandoned the traditional use of T-analysis;
3) Why the B-analysis, which is an appropriate method, should not be used alone, as was the case for many previous studies, but intead should be used in conjunction with the within-class data analysis (W-analysis).

Types of Data Analysis

The distinctions between the three types of analysis were discussed in studies by Bejar and Doyle (1974), Whitely and Doyle (1976, 1978, 1979), Doyle and Whitely (1974), and Sirotnix (1980). Their discussion convinced other researchers to accept the properties of the three types of analysis (B-analysis, W-analysis, and T-analysis). A brief description of each kind of analysis is drawn from them and is given as follows.

The B-analysis uses the instructor as the unit of analysis and provides results which are the most representative of teaching behavior (Bejar and Doyle, 1974), because the data from the B-analysis are the mean responses of students in each classroom to each item. Therefore, most individual bias and error will not be reflected in these means, nor will be the different individuals of different classrooms, nor how large the number of students in the class. It is true, according to Bejar and Doyle (1974), that this score would represent the "true" score of the instructor on teaching behavior. However, the students in the instructor's classes are not randomized into the classes; there is student self-selection into the classes. Therefore, there are still individual students in different classes, and, therefore, the W-analysis should be studied in combination with the B-analysis. The combination would help researchers in clarification of results. This type of study will reflect the student differences, as well as the instructor differences.

The W-analysis is an individual student-based approach. Correlations derived from it are based on individual differences where classroom differences are "controlled." This means that W-analysis correlation depends purely on individual differences in how various aspects of the

instructor behavior are perceived. Conceptually, one may consider that this is accomplished by computing correlation across individuals within each class separately, and then averaging these correlations across groups for a pooled within-analysis. Actually, the matrices of sums-of-squares and cross products are added element-by-element and this matrix is operated to from the pooled-within correlation matrix.

T-analysis uses the students as units of analysis and treats students from all classes as one group, regardless of instructor or section. T-analysis expresses a confounding of the two sources described above.

The studies of Bejar, Whitely and Doyle suggested to other researchers that T-analysis be abandoned and that the B-analysis is the kind of analysis which is the most appropriate to be used in the study of analysis of student ratings. However, Bejar, Whitely and Doyle have not suggested that the W-analysis should be combined in a study with the B-analysis. Wolins (in press) has done so.

Wolins (in press) suggested that the B-analysis should not be used alone, but only in combination with the W-analysis, and it would be profitable to perform factor analyses on both the between and within-class correlation matrices. Analyses of both of these require large sample groups. Wolins cited two studies to show that the T-analysis is inappropriate (Kennedy, 1975) and the B-analysis should not be studied alone (Brown, 1976).

Kennedy (1975), using the T-analysis, studied several hundred students of multiple sections (15 instructors) of a single course (an English course). He used the students as the unit of analysis, ignoring both

sections and instructors. He reported that students receiving the highest grades in class gave higher ratings to instructors. Wolins (in press) questioned these results, suggesting that the results were ambiguous as indicated by the previous discussion here.

Brown (1976), using sections as the unit of analysis, studied student grade and student ratings and reported that they had a positive relationship. Wolins (in press) questioned these results; they could be due to the individual variation that might randomly occur between classrooms, or they could be attributed to different instructors or sections. Wolins lacked confidence in the interpretation of Brown's study, since Brown had not shown that the W-analysis would have results of the same magnitude as the B-analysis. Wolins (in press) felt that if Brown had found the results of a W-analysis to be of lesser magnitude than the B-analysis results that he reported, then one could safely infer that, indeed, instructors who grade more liberally are evaluated higher by students.

It is suggested that it is necessary to analyze both B-analysis and W-analysis, rather than the B-analysis alone, as many previous researchers have done (Wolins, in press).

According to Wolins, B-analysis is appropriate to use, but one cannot have confidence in the results without support from W-analysis. Use of both analyses provides a basis for a clearer interpretation of the overall results of student ratings. Both B- and W-analysis can give an accurate explanation of results whether the results are due to the differences among instructors in classrooms or differences among students in each classroom. Wolins gave an example regarding the correlation between

the student of an instructor and grade given to the student. It is possible that the instructors gave the same evaluation on the average (grade according to student's performance in the course), but individual differences within classrooms produce the between-classroom variability only because there are different individuals in each classroom and the correlations between the classroom means on those two variables merely reflect these individual differences.

The above review was presented to explain why T-analysis was not used for this study, as well as to present why B-analysis was not used alone. This review was taken to be the first review with the purpose of providing the understanding of the types of analysis in this study before going into the other reviews of research.

The next part of the review is the review of previous studies which have studied the factors of teaching effectiveness.

Factors of Teaching Effectiveness

Identifying factors of teaching effectiveness is ancillary to the purpose of the present research. However, in the analysis of student rating data, the dimensions of teaching effectiveness appear as a by-product. A brief review of previous studies in this area follows.

Since students have been consumers of the teaching process, their judgments are often used as a source for identifying teaching effectiveness. Researchers can either analyze student rating data by T-analysis or by B-analysis. However, the studies that used the student as a unit of analysis (T-analysis) were cited by Bejar and Doyle (1974) and Whitely and Doyle (1976, 1978). Whitely and Doyle (1978) found that factors

from T-analysis and also W-analysis were factors of individual perceptions, which are implicit in students rather than instructor behaviors. They also found that T-analysis factors are likely to reflect largely implicit theories, while W-analysis factors are likely to reflect less (Whitely and Doyle, 1976).

It may seem that the present review of factors of teaching effectiveness would best be done separately from the studies that used the T-analysis and B-analysis; however, that is not possible. It was shown in the previous studies that even when the same type of analysis was used the different factors were found differently by those studies. The fact that different factors were found in different studies is not only due to the type of analysis used, but also to the particular questionnaire items of each study, the student or instructors sampled, as well as the course to which the sample belonged. The meaning of teaching effectiveness varies according to which courses were sampled (Desphande et al., 1970; Pohlmann, 1975b). Pohlmann (1975b) gave examples, including the study of Isaacson et al. (1963), which found an effective teacher in introductory psychology to be a "cultural," "artistic individual," while Desphande et al. (1970) found that effective engineering instructors receive high ratings on "structure" factors, factors of motivation, content mastery, and interrelational skills. Solomon (1966) reported that instructors in one area, for example social science, differed from their counterparts in other areas with respect to certain behavioral factors, such as clarity and permissiveness. There is no reason to expect the behaviors judged "successful" or "effective" in one content area of instruction to be equally so in others (Desphande et al., 1970). Therefore, when a study in this area uses a

particular group of courses and particular questionnaire items, one must understand that the factors of teaching effectiveness that emerge depend upon where the samples were drawn and the questionnaire items used.

There is a study which analyzed 1,279 classes at a midwestern university to identify factors of teaching effectiveness (Pohlmann, 1975b). Pohlman deviated from the tradition of many previous researchers who preferred using a certain course of study by using varied courses. He found factors such as preparedness and organization in presenting course material, achievement of the course objectives and increase in students' appreciation of the subject matter to be important facets of teaching effectiveness. Costin et al. (1971) reviewed many studies and concluded that the following factors were reported in many studies: preparation, clarity and stimulation of students' intellectual curiosity. Factors such as organization or structure, teaching skill, communication or overall fluency, and student rapport were the factors mainly found (Linn et al., 1975). Feldman (1976) had summarized the results of nine different studies in which students were asked to describe the characteristics of the "best" teacher without being given any specific format for their answer. He reported that the traits listed as most important by students were (in order):

- concern or respect for student (including friendliness);
- knowledge of subject matter;
- stimulation of students' interest;
- availability and helpfulness;
- encouragement of questions and discussion;

- ability to explain clearly;
- enthusiasm for the subject or for teaching;
- impartiality;
- preparation for (and organization of) the course;
- instructionary skills.

Wotruba and Wright (1975) summarized twenty-one selected research studies in which various groups had been asked to identify the quality of teaching effectiveness. They reported the following factors (in order):

- communication skill -- interprets abstract ideas and theories;
- favorable attitude toward students;
- knowledge of subject;
- good organization of subject matter and course;
- enthusiasm about subject;
- fairness in examination and grading;
- willingness to experiment -- flexible;
- encouragement of students to think for themselves;
- interesting lecture.

Visual inspection of the above teaching effectiveness factors suggests that factors found so far are similar across the various studies. Direct comparisons of factors could not be done, due to differences in rating items and factoring techniques, as well as to other problems discussed earlier.

Inconsistent Findings of Student Ratings and Student Characteristics

The study of identifying factors of teaching effectiveness is not the only interesting aspect of a study of student ratings. Other studies of student ratings, such as the generalizability and validity of student ratings, have been studied by many researchers. The study of generalizability and validity focus on the quality of student rating information. The generalizability is representative of whose opinions and observations the data reflect, how well the sample portrays the totality of the person's teaching, and to what extent situational factors influence the evaluation (Doyle, 1975). Validity provides meaning which draws from data of the ratings in both sense of internal structure (i.e., item intercorrelations and dimensions) and relationship of external criteria and other measures of student accomplishments.

Student rating studies of generalizability and validity have been inconsistent (Costin et al., 1971; Feldman, 1977; Graham, 1973; Aleamoni and Hexner, 1980). Many explanations have been given for this inconsistency. Graham (1973) cited the limitations of many previous research studies (see page 23 of her study), which she felt produced the disagreement which exists concerning the effects of each of the variables on student ratings. Graham also cited the methodology of previous studies which used the univariate method and said that using multivariable analysis of variance (MANOVA) is appropriate.

Whitely and Doyle (1979) listed the three failures of previous research which may explain the inconsistent findings of previous studies.

They are: (1) failure to study the validity and generalizability variables simultaneously and in a single paradigm, (2) failure to distinguish among the different types of analysis (T-analysis, B-analysis and W-analysis), and (3) failure to distinguish among different types of instructors. The case of the second failure is often seen in studies which favored using only one type of analysis (a single type, either T-analysis or B-analysis). The third failure is often found in many previous studies that have favored the use of certain courses to study and others use multiple sections, but yet these sections come from a single course. This is a failure to distinguish among different types of instructors.

Recently, various investigators have incorporated methodological improvement, which has provided more definitive results (Brandenburge et al., 1977; Smith, 1979a). For example, Aleamoni and Graham (1974) advocated a multiple validational approach that permits the examination of several potential biasing factors in combination. In addition, other researchers (e.g., Bauswell et al., 1975; Doyle, 1975; Morsh et al., 1956) have recognized the class rating means rather than individual ratings as the proper unit of analysis (Bauswell and Bauswell, 1979). Smith (1979a) reported that reliability and validity of student ratings have been examined extensively using the classical measurement theory (for a review, see Doyle, 1975), but recently, generalizabilities theory (Cronbach et. al., 1972) has been applied to this area of student ratings instead. The generalizability theory, applied to use in study of student ratings, has been utilized by Kane et al. (1974, 1976) and Kane and Brennan (1977). The class is

used as a unit of analysis and when it is being analyzed, it is in a condition of observation characterized by item facets and student facets. It was suggested (Gillmore et al., 1978) that, in order to obtain high generalizability coefficients, data should be correlated from as many courses as possible. This is another area where it was often found that previous studies had used only a course to study. Another important finding from study by Gillmore et al. (1978) was that they found that instructors, rather than course offerings, affect the student ratings. The individual from the studies of Hogan (1973), Bauswell et al. (1975), and Feldman (1978) supported the findings from the study of Gillmore et al. (1978), who reported that the variance component associated with instructors is relatively large, while that associated with the course is negligible. Bauswell et al. (1975) and Hogan (1973), who investigated the correlation of instructor and course in three different situations, studied (1) same instructor teaching same course (SI-SC); (2) same instructor teaching different courses (SI-DC); and (3) different instructors teaching the same course (DI-SC). Bauswell et al. (1975) found the following correlations: SI-SC = 0.69; SI-DC = 0.33; and DI-SC = 0.17. Hogan (1973) found SI-SC = 0.70, SI-DC = 0.40 and DI-SC = 0.19. In the review of the studies of these three sets by Feldman (1978) it was found that the three sets' correlations varied as follows: SI-SC = 0.62 to 0.80, SI-DC = 0.29 to 0.54, and DI-SC = 0.04 to 0.20. The correlation of SI-SC and DI-SC is the index of the importance of the instructors. The low correlation of DI-SC clearly shows that the instructor affects the student rating more than does the course. The SI-DC set is an index of the importance of the course content. In any case, it could be concluded that legitimacy of

using the student evaluation as a basis of measuring an instructor's effectiveness is based on the assumption that class effect is due mainly to the instructor's effectiveness.

The above review was intended to show that previous research based on sample groups from only one course or multiple sections of a single course is not appropriate to use in analyzing student ratings because it will be too limited or would not ensure valid results. A small sample group is not suitable for the analysis of student ratings because it is necessary to analyze both the B-analysis and the W-analysis. Therefore, the study by Whitely and Doyle (1979), which was a study of generalizability and validity of student ratings from B-class and W-class data, could be considered a failure due to the small sample group used. Their sample groups were too small (11 teaching assistants of a recitation class and 5 professors), and they studied only one course (mathematics). Their study reported that neither within-class nor between-class student ratings were related to student background characteristics, ability, year in school, sex, or whether the course was required. These results are questionable due to inappropriate sample groups. Their study is also limited to study of lower student level. Even though student level was not reported, it was probably predominantly freshmen (in a beginning mathematics course). The Whitely and Doyle (1979) study is not the only study that had some limitations. Other studies (using B-analysis) were similar (e.g., Doyle and Whitely, 1974, using beginning French course; Bendig, 1954, 11 instructors of multisections of an introductory psychology course; Frey, 1973, 8 instructors of multisections of a multidimensional calculus).

In the opinion of this researcher, it is more suitable to study multiple sections of a variety of courses and analyze both by B-analysis and W-analysis to yield better results, especially in the area of validity of student ratings which have been studied and found to be inconsistent. B-analysis could give clear results about instructors when a variety of instructors from a variety of courses are studied. It is important that this variability is present because course effects exist (Feldman, 1977). Feldman (1977), in comparing student ratings of different academic areas, found that student ratings of English, humanities, art and language, fell mostly in the high and medium range. Social sciences tended to be in the medium or lower third of rankings. Other fields of sciences, mathematics, and engineering were usually in the lower two-thirds of the rankings. However, this report gave only information about placement of academic fields relative to one another; it did not show that instructors in a certain field are more likely than another to receive absolutely high (or low) ratings.

Student Characteristics and Student Ratings

Instructor effect and course effect are not the only two variables which affect validity of student ratings. Student characteristics are another variable that affects student ratings. The studies of correlation of student characteristics and student ratings have reported inconsistent results (Costin et al., 1971; Kulik and McKeachie, 1975). It has been found that sometimes positive, sometimes negative, and sometimes zero correlations exist between student ratings and various student characteristics, such as sex, student grade, etc., as well as required vs. elective

courses. Females tend to rate the instructors higher than male students. Students taking a course as an elective are believed to give higher ratings than do students in required courses. Regarding student grade, one widespread belief stems from a grading leniency basis -- easy graders might receive better student ratings than hard graders.

Following are separate reviews of studies relating to student gender, required vs. elective course, and student grades to student ratings. These are the purposes of the present study. The researcher was unable to find any literature regarding how student gender or required vs. elective courses relate to student ratings. The review includes previous research which has been done by B-analysis. Research which used T-analysis will be mentioned only when it is necessary to do so.

Student gender and student ratings

Conflicting results of student sex and student ratings have been obtained. Gender has been found to affect the student ratings with some (but not complete) consistency across studies. Feldman (1977) reviewed nearly 50 studies which used either T-analysis or B-analysis. Half of those studies in his review reported no correlation between student gender and student ratings, while the other half of those reported that there is a correlation which is statistically significant. A selection of findings in this area which used only the B-analysis is presented as follows.

McKeachie et al. (1971), who analyzed several classes, found that female students gave higher ratings to instructors on the factor of "structure." Female students tended to rate instructors higher on the factor of specific objectives of the course, but not on other factors (Elmore and

LaPointe (1975). An analysis of 87 instructors by Rader (1968) revealed that student ratings were not substantially related to student's sex.

Required course vs. elective course and student ratings

Course variables commonly investigated as possible causes of invalidity in student ratings are the required vs. elective nature of the course. Several investigators have found that students who are required to take a course tend to rate the instructor of the course lower than students who elect to take it (Pohlmann, 1975a; Gillmore and Brandenburge, 1974; Brandenburge et al., 1977). Brandenburge et al. (1977) analyzed over 3,000 class sections over two semesters and found that the proportion of students in a class taking the course as an elective course was positively related to average ratings even when controlling for the average expected grade of students in the class, class size, the course level, student gender, and rank of instructors. In contrast, Whitely and Doyle (1979) reported that neither within classes nor between classes were student ratings related to whether the course was required or not.

In conclusion, Feldman (1978) reviewed a number of studies which used either T-analysis or B-analysis. The results did not seem to depend on the kind of analysis used. Feldman (1978) stressed that while some found particular relations of student ratings and required vs. elective course, this was not always apparent in other studies, and in studies which did find a correlation, the correlation was generally small. When the "average" of students taking the class as elective appeared in studies, the relationship to their ratings is found generally positive and of small to moderate strength.

Expected grade (and actual grade), GPA, and student ratings

Grades are often assumed and considered to be one of the greatest potential sources of invalidity of student ratings. There is a potential indicator that student ratings could be invalid since students will merely rate instructors in accordance with the grade received or expected in a particular course. The relationship of student grades and student ratings has been found to be inconsistent as the following will review.

There are four important reviews of student grade and rating relationship which have been done by Graham (1973), Aleamoni and Hexner (1980), Feldman (1976), and Stumpf and Freedman (1979). The present study intends to review this area on the basis of these four studies which had never a great number of research studies appearing in their reviews.

Research which found that there was no relationship between grade and student rating or the relationship was near zero (report of 20 studies reviewed by Graham (1973, p. 14), 20 studies by Aleamoni and Hexner (1980)). Some of these studies appeared twice in the review by Graham (1973) and Aleamoni and Hexner (1980), but it could be said that there are about 20 researchers who found no relationship or where relationships were near zero. In contrast, there are nine reviews by Graham (1973), 25 studies by Stumpf and Freedman (1979), 28 studies by Aleamoni and Hexner (1980) that found the positive correlation of student grades and student ratings. Again, there are some studies which appear in more than one review, but it could be said that nearly 25 studies of those reviewed found a positive correlation.

Holmes (1972) not only found that there is no significant relationship between student grade and student ratings, but also found students who received low grades tended to rate the instructors significantly lower in factors of organization, preparation of materials and examinations. Centra and Lin (1976) found the same; the lower-graded student with less interest in the course tended to be more critical of examinations or coursework.

Some interesting explanations for the inconsistent findings among the research were given by Lolli (1977) and Stumpf and Freedman (1979). Lolli (1977) pointed out that the inconsistency was dependent upon the data collection methods. Some studies have examined cases in which grades were awarded prior to the students' evaluation of their instructor. He gave examples such as Remmers' (1960) study, and studies by Bendig (1953), Russell and Bendig (1953), Anikeef (1953), and Brown (1976), while other studies required students to evaluate their instructors prior to the awarding of grades (e.g., Garverick and Carter, 1962; Weaver, 1960). Lolli has found that even when they collected data in different cases, findings were the same. For instance, Garverick and Carter (1962) and Russell and Bendig (1953) found positive correlations even when data was collected in different cases as Lolli (1977) suggested. Therefore, it seems as though there are other things that cause the inconsistencies. Stumpf and Freedman (1979), as well as Feldman (1976), reviewed a great number of research studies by grouping the research into two tables. One is of the research which used the student as the unit of analysis (analysis of total-class data) and the other is the research which used the class (or instructor) (between-class data), as the unit of analysis. Feldman (1976) reviewed

over 40 studies and presented four interesting tables. In his Table 1[1] where he cited 15 studies, he showed the summary of studies giving between grades (and GPA) and courses or teacher evaluation, individual college student as unit of analysis, with data pooled across two or more classes. His Table 2[2] (containing 16 studies) is the summary of studies relating grades to course or teacher evaluation-comparison of students in different grade categories, with data pooled across two or more classes. The studies reviewed in Table 1 were the analyses which were done on data that had been pooled (unweighted), gave the correlation coefficients, while the studies in his Table 2 used the other techniques of data analysis.

The conclusion of studies in Table 1 showed that there is a positive correlation and most are statistically significant. These correlations range in size from 0.10 to 0.46. From Table 2, most studies gave the estimated tendency of student expected grade and their ratings, and no strength of association was shown. Feldman (1976) assumed from his summaries of this group of studies, that the range in strengths of association of those studies is not different from those results of studies in Table 1. In his Table 3,[3] summaries of a study of a psychology course taught by 3 different instructors, data are analyzed separately (like study of within-class analysis). The correlation of each instructor and the ratings are shown. The last group in his Table 4[4] is the group of

[1]Table 1, Feldman (1976, p. 71).

[2]Table 2, Feldman (1976, p. 78).

[3]Table 3, Feldman (1976, p. 87).

[4]Table 4, Feldman (1976, p. 95).

studies which used class as unit of analysis. They were still tending toward a positive correlation between student grades and student ratings, although the results of this group of studies were more variable than those which used students as the unit of analysis. Feldman criticized the studies in pooling data across classes that those studies (in his Table 1) might mask the fact that grades are more strongly associated with evaluation in some classes, while only weakly correlated, or not at all, in other classes (and possibly even negatively related). This made the results more, rather than less, strongly associated with student ratings. Also, results possibly came out from the results of the interaction of student attitude and value as well as the bias from student gender. He made these remarks to those studies because it was shown in a study that he presented in Table 3, which analyzed correlation of factor ratings and student ratings to each instructor of a course (such as the study of within-class data), which was done by Yonge and Sassenrath (1968). Their separate analyses for each instructor revealed how a grade is clearly associated with each instructor. The rating of one of the three instructors was influenced by grades, and rating was correlated highly on every factor, while it was not for the other instructors. This pattern also was found in the study by Weigel et al. (1971) with seven classes; Doyle (1972) with six classes, and Holmes (1971) with seven classes. All of these studies showed moderate association in each study that grade is either less strongly associated or not at all associated with rating in other classes. These results made Feldman (1976) suggest that data should not routinely be pooled across classes without checking on the advisability

of doing so. Moreover, his assumption made further variation in outcomes across nonpooled data (in existing research) meaningful and not merely haphazard; the important task is to find out the conditions under which grades in a class can be expected to be positively associated with student evaluations, as well as the circumstances under which such associations are especially strong. Unfortunately, these conditions and circumstances have not yet been established. Feldman (1976), in his final conclusion for this review, stated that it cannot be said that grades tend to bias student ratings, nor can it be concluded that they do not. Brown (1976) found grading bias to be dependent on how valid the grades are in each section of the course.

Stumpf and Freedman (1979), did the same as Feldman (1976), but also conducted a study and analyzed data on both total-class data and between-class data across data of three academic terms. They reported that the inconsistency of student grades and student ratings of the previous research came from the methodology differences among previous research that didn't distinguish between individual and class effects. They presented in Table 1[1] (25 studies) the group of studies which used individual as unit of analysis, and in Table 2[2] (14 studies), the group of studies which used the class as the unit of analysis. Twenty-five studies shown in their review Table 1 had total students of 30,000 and 1,500 classes. The fourteen studies (shown in Table 2) had total classes of 7,700. They computed this total from Tables 1 and 2, and found that median correlation of

---

[1]Table 1, Stumpf and Freedman (1979, p. 294).

[2]Table 2, Stumpf and Freedman (1979, p. 296).

instructor rating with actual or expected grades at class level (between-class analysis) is larger than at the individual level (total-class analysis; 0.37 vs. 0.18). The range at the class level is also greater (-0.75 to 0.75 vs. -0.2 to 0.85). Their study, which analyzed the total-class data (7,893 students) and between-class data (297 classes), showed that the total-class analysis student ratings was significantly related to expected grades ($p \leq 0.55$), although the effects are not large (median $r = 0.22$); in comparison to the results for the between class data, the correlation of student ratings with student grades are generally significant or larger (median $r = 0.39$). Their study showed that the subject matter effect (i.e., the degree to which students like the subject) has the strongest relationship with student grades. It was also found that between-class differences in grade substantially contribute to the variances in the total sample. The between-class covariance is of sufficient magnitude to be of concern in itself (14.5%), whereas the total-class correlation is not large enough to suggest a serious bias issue. There is a concern that the administrator would have to consider whether the results come from total-class or between-class data. That is, when the analysis of total-class data was reported, the size of grade-rating covariance does not appear to be a serious bias threat. When the analysis from the between-class data was reported, the covariance of ratings with grade could be viewed as undesirable by some faculty because of the substantial differential advantage occurring to those expected to be easier graders. When decision-making is for pay, promotion and tenure, the results from the analysis of between-class data could give a clearer picture of the instructor. From a

practical point of view, the importance of separating and quantifying individual class effects largely depends on the unit of analysis used for feedback and performance appraisal purposes (Stumpf and Freedman, 1979). Howard and Maxwell (1980), as well as Brown (1976), are in agreement that many of the inconsistencies in the literature of this area are due to methodological dissimilarities among studies. Many of the investigations found weak relationships between grade and student rating employed the data of individual students as the unit of analysis (Howard and Maxwell, 1980). Howard and Maxwell (1980) stated that the inconsistencies in the literature vanish rapidly if one considers only class mean as unit of analysis; further, the mean rating approach is also preferred, since this is the unit used for the important practical purposes of student rating (e.g., for promotion, tenure, salary).

From the above review, there is a study by Whitely and Doyle (1979), which was reviewed by Stumpf and Freedman (1979); it focuses research done on both between- and within-class analysis. Whitely and Doyle (1979) found covariation of student grades and rating of 0.35 when class is unit of analysis (between analysis) and found -0.02 in the within analysis when students were used as the unit of analysis. Whitely and Doyle (1979) reported that neither within-class nor between-class data were student ratings in their study relative to the students' background characteristics and ability in school, sex, or whether the course was required or elective.

The present research questions their findings and the research also cited earlier in this chapter.

Although their study seems appropriate, the sample is not appropriate; the study of the between-class data and within-class data requires a big sample which will provide the acceptable results. It is, however, the only study which analyzed both between- and within-class data. No other studies were reported in the review by Feldman (1976) or Stumpf and Freedman (1979) which had other research done on both between- and within-class data analysis.

Different studies of this area have found that some factors showed the correlation of student ratings and student grade, but not the other factors. It is hard to make a conclusive review of what factors had associated to student ratings and student grade. However, there are common findings that were found to be the same in the three studies. Pohlmann (1975a) found the factor of course difficulty to be related to student ratings and student grades. Holmes (1972), who reported no significant correlation between expected grade and student ratings, did, however, find that students who thought they received a lower grade tended to rate the instructor significantly lower in factors of organization, preparation of materials, and the examinations. Centra and Lin (1976) found that higher grades/expected grade, high GPA, lower class level, tended to rate instructors higher in factor of examinations, which was rated lower by students who had lower expected grades and they also tended to be more critical of examinations and course work.

The relationship of this area is different from study to study. Most of the reports which found the relationship also found that different factors related to student ratings. The above sample findings were more

relevant where others were not. It is possible to assume that, in this area, the relationship is dependent upon course and instructor. The majority of research reviewed by Stumpf and Freedman (1979), Feldman (1976), Aleamoni and Hexner (1980) did not study grade relationship with course and subject component; only the instructor criteria were comparable across unit of analysis. The relationship of interaction of different course and instructor could not exactly be the same.

Conclusion

The overall review in this chapter could be summarized as follows.

It appears that there are many inconsistent findings regarding student ratings. The different findings appearing in previous research are due, in some studies, to their limitations, and in others, to the inappropriate analysis method, as reviewed in the methodology chapter. They were reviewed for this research in order to avoid those limitations and that type of analysis method. In addition, there was no report from previous studies that the interaction of the variables and their associations which affect student ratings in total had ever been studied. This fact shows the need for more research and suggests that the research be conducted to cover the points not covered in previous studies. However, one has to keep in mind that the study of the variables that affect the student ratings of the class varies from class to class. It means that one should not compare the student ratings of one class to others. This is because the student ratings are not the result of teaching effectiveness only. The variables of students and type of course that contains students with variability effects the student ratings of a class differently. Therefore, the

results of student ratings of class should not be really compared, especially if they are different courses. Each course has its own contribution and effect on student ratings results, which are due to the instructor's teaching effectiveness and the variables of the students' characteristics and the type of course.

METHODOLOGY AND PROCEDURE

The review of literature in the previous chapter revealed that there exist many inconsistent findings in the studies of student ratings. It has been suggested that improvement in the method of analysis and that greater range of coverage in the sample groups are mandatory for any new undertaking aimed at clarifying these inconsistencies. Therefore, since the present ratings were obtained in areas where inconsistent results have been found, a combination of B-analysis and W-analysis is used with a larger sample to better represent the instructors and students of various disciplines.

Sample

The sample group is comprised of classes from multi-sectional courses offered by various departments at Iowa State University of Science and Technology during spring quarter, 1981. Each class met the following criteria:

1) Each must be a class of a multi-sectional course which is taught by different instructors in those sections;
2) If an instructor taught more than one section in the same course, his (or her) sections were combined and counted as one;
3) Each section must have a minimum enrollment of 25 students.

There were 88 classes (88 instructors) sampled, totaling 2,107 students, of which 1,228 were males and 884 were females. The students in those classes break down according to year standing as follows: 312 (15%) freshmen, 613 (29%) sophomores, 676 (32%) juniors, 492 (23%) seniors and 9 (1%) graduate students.

Instrument

The item questionnaire is included in the Appendix A. It was the instrument of this study, and it contained 65 items. The questionnaire items 1-55 were assembled from a variety of published student rating forms and from student rating forms of selected universities in the midwest and from commercial student rating forms. Some items were added by this researcher. Items 55-65 were used for student information.

The basis for selecting these 55 items from the great number available was judgmental by looking at student evaluation forms of many resources, as mentioned above. The primary consideration in making these judgments was to select several items defining each of the matrix factors defined by others. It was not possible to select items for all such factors, but it is hoped that the ones selected define those that have been commonly defined. A further consideration is the classroom time necessary to administer the questionnaire. If too long, it would result in little cooperation from the instructors. If it is too short, it would not cover the more commonly defined factors. The 55 items selected represent a compromise between these two considerations.

Analysis

The responses to each of the evaluation items were transformed to normal deviates (probit) as described by Wolins and Dickinson (1973). This 99-point scale and the associated transformation has provided useful results in a large number of studies involving affection stimuli (e.g., Hendricks, 1974; Elrod and Crase, 1980). The transformation changes the 99-point scale by separating more widely response

differences in the two ends of the scale and, relatively, compressing the center part of the scale.

As previously discussed, the intercorrelations among the items were obtained twice: between and within. This was done using the MANOVA part of the Statistical Analysis System (SAS). The factor extraction procedure (PRINIT) and the rotation procedure (VARIMAX) were also implemented by SAS.

Following extraction, the number of factors rotated was based on the interpretability of the results. The score (Cattell, 1978) was inspected and, on this basis, the maximum number of factors was rotated. These larger solutions were inspected and discarded when the factor with the smallest eigenvalue did not make sense. The solution chosen was the largest one where each factor was judged to be interpretable.

Following this, the rotated solutions were compared, judgmentally. On the basis of these analyses, scores were derived from the items which loaded highly and uniquely for each of the between factors, where possible. Some factors, though meaningful, were not measurable because the items that loaded on them loaded higher elsewhere and the loadings on these factors were not large.

The scores resulting from the factor analyses were correlated with student characteristics (sex, grades, etc.). As before, both a between- and within-correlation matrix was obtained.

Two analyses of variance were done. The first recognized only between and within classroom variability and utilized the whole data set. The second recognized sex as a repeated measure of classrooms and sections

and used only those classrooms that had at least five students of each sex. The latter analysis did not use individual scores. The error term for courses was sections within courses and the error term for sex and the course interaction with sex was the sex by course interaction. This analysis was done on the unweighted means. Any questionnaires with incomplete responses were thrown out.

Finally, pictorial representations of the data were constructed as aids in interpretation.

RESULTS AND DISCUSSION

There are two parts in this chapter which report findings. The first part presents the factor analysis of the questionnaire items. Three tables are presented in this part. Table 1 shows the between factor loading of items 1-55, in which 10 factors were found. Table 2 presents the within factor loading of items 1-55, which are distributed in ten factors. Finally, items and factors in both tables were screened and selection was made of only the items that had highly relative loading in factors. Only 42 items out of the 55 items and eight factors in the between analysis and nine factors in the within analysis were selected, and these are presented in Table 3. This table shows the item loading in each factor. On the left of this table is the loading of items in the between classroom factors, and on the right is the loading of items in the within classroom factors. From Table 3, the exact phasing of items in the eight factors is shown in Appendix B. All details of findings and discussion of the factor analysis of the questionnaire are given in detail in the first part of this chapter.

In the second part of this chapter, the analysis of the correlation and variance of student rating factors are presented. The analysis of the correlation of the factor scores and student characteristics is shown in Table 4 and the analysis of variance is shown in Table 5. Table 4 presents both within and between correlations of the factor scores and student characteristics.

Table 6 presents the within group correlations below the diagonal, as in Table 4. In addition, the diagonal elements contain the F-ratios of the between- to the within-group variance. Above the diagonal are

indices constructed by pre- and post-multiplying the between correlation matrix by the diagonal matrix composed of the square root of these F-ratios. These between indices are expected to be the same as the within ratios. These between indices are expected to be the same as the within correlations under null conditions. Since classrooms differ significantly in all variables, these between indices are uniformly larger (in absolute value) than the within correlations.

Thus, the between matrix is substantially different from its expectation, the within matrix in all respects. On this basis, one can observe that the apparent resemblance of the between to the within correlation matrix is somewhat misleading in that these two matrices are not scaled the same.

The interpretation from the results based on Table 4 is given in detail. Also, the study of analysis of variance of each of the eight factor scores is done with the recognizing as sources of variability of course, sections within courses, sex of student, and course by sex interaction. This is shown in Table 5. The overall interpretation of the second part of this chapter is based upon the results in Tables 4 and 5. The last part of this is the conclusion of the chapter.

Factor Analysis of the Student Evaluation of Instruction Questionnaire

The items used to measure a particular factor derived from the factor analyses of the between- and within-classroom correlation matrices were selected as follows:

1) The item must load relatively high on the factor on which it loads highest;
2) It cannot load highly on any other factor except when all the items loading highest on a particular factor load some particular other factor. That is, correlations among factors should be avoided when possible.

The items that meet these two criteria are listed in Table 3. The implementation of the selection procedure is illustrated by items 30 and 31. These two items load on both between- and within-analyses. For the within analysis, which is presented on the right half of the table, these two items are not used to measure Factor II because they load substantially on Factor I also. However, these items are used to measure the same factor (apparently) from the between analysis because Factor I from the within analysis does not occur for the between analysis, and, as a result, these two items have no high loadings other than the highest one, which is in the between analysis of Factor III.

In Table 3, indicated in parentheses, are factor loadings from one or the other analysis that are the highest loading for the item for one analysis but not the other. For example, for both analyses, items 16 and 17 load on the same factor. However, item 40 loads heavily for the between analysis (-0.64), but substantially less than for the within analysis (-20). Such results could occur due to sampling fluctuations, but an attempt will be made to interpret such results.

The following presents the details of each factor and shows exact phasing of items in each factor in Appendix B.

Table 1. Between factor loading

	Communality	Rotated factors loading matrix (denormalized)									
		I	II	III	IV	V	VI	VII	VIII	IX	X
1	.83	.05	-.15	-.25	.85	.04	-.09	.02	.04	-.07	.07
2	.77	.18	-.23	-.67	.37	.07	-.01	-.12	.17	.14	.02
3	.73	-.31	.30	.31	-.17	-.59	.10	.07	.02	-.07	-.24
4	.83	.50	-.61	-.25	.18	.15	-.09	-.25	.03	-.04	.15
5	.87	.77	-.37	-.25	.10	-.02	-.17	.02	.00	-.17	.07
6	.89	.66	-.51	-.24	.08	-.03	-.14	-.18	.16	.12	.22
7	.73	-.30	.18	-.01	-.11	-.19	.74	-.02	-.11	-.06	-.03
8	.46	.02	-.06	-.11	.03	.33	.13	-.49	-.17	.05	.23
9	.75	.34	-.17	-.32	-.10	-.12	.02	-.45	.52	-.05	.10
10	.72	.57	-.34	-.05	.08	.22	-.32	-.01	.32	-.08	-.10
11	.93	.87	-.19	-.17	.04	.17	-.18	-.04	.05	-.19	-.08
12	.82	.80	-.06	-.06	.15	.27	-.18	.05	.01	-.15	-.16
13	.75	.42	-.51	-.18	.39	.16	-.05	-.27	.01	-.08	-.13
14	.92	.56	-.56	-.15	.10	.20	-.27	-.25	.05	-.17	.20
15	.94	.53	-.64	-.12	.04	.17	-.25	-.22	.08	-.21	.20
16	.81	.21	-.46	-.06	.15	.68	-.07	.21	.01	.11	-.07
17	.88	.24	-.46	-.16	.11	.73	-.21	-.03	.03	-.02	-.04
18	.88	.37	-.35	-.67	.15	.30	-.16	-.14	.02	.02	.12
19	.85	-.23	.09	.79	-.16	-.18	.03	.30	-.05	-.11	-.05
20	.84	.16	-.86	-.10	.13	.14	-.13	.03	.09	.06	-.00
21	.93	.75	-.45	-.18	.13	.11	-.23	.05	.12	-.07	.15
22	.86	.37	-.70	-.29	.14	.20	-.18	.01	.19	-.09	.13
23	.80	.15	-.46	-.01	.19	.19	-.69	.03	.07	-.09	.01
24	.88	.54	-.56	-.23	.11	.14	-.33	-.08	.04	-.18	.20
25	.86	.63	-.40	-.21	.05	.21	-.26	-.04	.17	-.07	.33

26	.88	.32	-.40	-.12	.26	.30	-.66	-.03	.09	.03	.02
27	.91	.41	-.58	-.28	.29	.31	-.20	-.11	.26	-.02	.10
28	.86	.83	-.20	-.22	.07	-.02	-.03	-.20	.07	.18	-.05
29	.91	.73	-.39	-.32	.07	.02	-.05	.07	.13	.12	.28
30	.95	.95	-.33	-.63	.18	.28	-.13	-.24	.08	.04	.12
31	.94	.94	-.33	-.68	.19	.16	-.13	-.29	.12	.01	.05
32	.69	.31	-.34	-.22	.02	.14	-.17	-.22	.39	-.40	-.15
33	.44	-.13	-.07	-.06	-.11	-.14	.05	-.60	.08	-.03	-.08
34	.36	.18	-.10	-.25	.04	.13	.04	-.46	.15	.04	-.00
35	.88	.21	-.84	-.17	.13	.13	-.08	-.06	.12	.09	-.17
36	.85	.70	-.19	-.19	.36	.09	-.04	-.13	.05	.35	.07
37	.93	.29	-.28	-.16	.85	.06	-.09	.03	-.09	.06	.06
38	.66	.52	-.46	-.21	.14	.17	.15	.09	.16	-.16	-.01
39	.60	.53	-.29	-.11	.18	-.01	.13	-.06	.36	-.20	-.00
40	.69	-.22	.16	.03	.02	-.64	.42	.07	-.07	.14	.00
41	.87	.80	-.51	-.20	.13	.13	-.08	-.12	.15	.29	-.07
42	.94	.32	-.80	-.20	.20	.20	-.14	-.09	.18	.08	.05
43	.78	.22	-.00	-.45	.06	-.39	.11	-.47	.26	-.26	-.01
44	.88	.26	-.40	-.73	.11	.05	-.21	-.23	.02	-.11	-.06
45	.56	.44	-.26	-.27	.16	.14	.26	-.14	.03	.00	.31
46	.58	.06	.00	-.52	-.06	-.14	.08	-.38	-.04	-.33	.11
47	.95	.46	-.67	-.37	.20	.14	-.20	-.18	.01	-.15	.04
48	.83	.42	-.67	-.27	.13	.08	-.26	-.12	-.01	-.01	.20
49	.96	.88	-.20	-.21	.08	.16	-.22	.01	.04	-.11	-.08
50	.87	.58	-.45	-.52	.14	.05	-.02	-.05	.07	-.04	.19
51	.58	-.11	.22	.07	.20	-.02	.20	.06	-.65	-.04	-.03
52	.85	-.21	.22	.21	-.77	-.07	.22	-.05	.22	-.04	.12
53	.91	.72	-.42	-.22	.12	.10	-.11	.04	.08	.06	.35

Table 1. *Continued*

Communality		Rotated factors loading matrix (denormalized)									
		I	II	III	IV	V	VI	VII	VIII	IX	X
54	.95	.55	-.66	-.27	.26	.18	-.13	-.12	.06	-.04	.03
55	.44	.11	-.17	-.56	.19	-.04	.06	.07	.03	-.10	-.01
56	.70	.39	-.10	-.45	.34	.17	.06	.28	.29	.15	-.07
Percent removed (total variance) by each of 10 factors											
		22.70	17.43	10.84	6.32	5.87	5.35	4.30	3.16	1.95	1.93
Percent total variance removed by 10 factors											
	79.84										

Table 2. Within-factor loading

Communality	Related factors loading matrix (denormalized)									
	I	II	III	IV	V	VI	VII	VIII	IX	X
1 .26	-.07	.19	.05	.12	.03	.09	.14	-.09	.40	-.04
2 .44	-.19	.48	.09	.14	-.19	.17	-.01	.07	.27	.07
3 .29	.18	.46	-.07	.06	.46	-.08	-.01	.07	-.06	-.01
4 .48	-.50	-.15	.15	.04	-.17	.17	.04	.05	.30	.14
5 .59	-.56	-.15	.42	.05	-.13	.11	.00	.01	.20	.10
6 .56	-.57	-.30	.33	-.03	-.06	.08	.01	.03	.14	.07
7 .38	.09	.08	-.09	-.08	.50	-.02	-.32	-.02	-.01	.02
8 .24	-.04	-.16	.04	-.01	-.12	.42	.04	-.13	.04	.02
9 .30	-.18	-.20	.08	.12	.01	.45	.07	-.02	.02	.01
10 .45	-.28	-.16	.28	.10	-.27	.36	.22	-.00	.09	.02
11 .72	-.27	-.09	.78	.05	-.13	.08	.05	-.02	.04	-.02
12 .43	-.25	.06	.55	.02	-.17	.10	.05	-.08	.07	.02
13 .29	-.31	-.08	.13	.06	-.14	.30	-.00	.01	.23	.10
14 .74	-.70	-.31	.20	-.07	-.16	-.04	.01	-.05	.09	-.27
15 .76	-.75	-.25	.18	-.06	-.07	-.04	.02	-.08	.08	.29
16 .54	-.19	-.16	.14	.16	-.17	.13	.06	-.61	.12	.01
17 .59	-.20	-.26	.11	.08	-.31	.18	.14	-.55	.10	.01
18 .62	-.24	-.67	.11	-.00	-.16	.17	.06	-.18	.09	.04
19 .50	.12	.63	-.04	-.12	.25	-.05	.07	-.00	-.03	-.07
20 .55	-.62	-.04	.06	.32	-.09	.16	.03	-.10	.07	.05
21 .58	-.64	-.13	.28	.13	-.08	.12	.08	-.15	-.01	.12
22 .49	-.60	-.14	.15	.14	-.14	.15	.11	-.06	.09	.07
23 .66	-.33	-.12	.11	.11	-.26	.16	.60	-.09	.21	.06
24 .51	-.62	-.13	.20	-.02	-.01	.09	.20	-.08	.03	-.11
25 .56	-.59	-.29	.28	-.09	-.04	.03	.17	-.08	-.02	-.02

26	.71	-.30	-.15	.12	.12	-.38	.18	.60	-.11	.16	.03
27	.51	-.48	-.26	.11	.22	-.19	.21	.23	-.08	.10	.05
28	.42	-.41	-.10	.33	.12	.01	.16	.17	-.06	-.06	.23
29	.64	-.60	-.18	.24	.21	-.01	.04	.12	-.12	-.03	.34
30	.73	-.42	-.69	.15	-.06	-.12	.02	.15	-.12	.05	.08
31	.73	-.40	-.70	.10	.09	-.08	.10	.14	-.08	.08	.05
32	.43	-.38	-.35	.12	.26	-.08	.18	.17	-.10	.10	.01
33	.18	-.06	-.22	.09	.09	.12	.20	.18	-.09	.06	-.10
34	.19	-.09	-.06	.07	.30	.02	.27	.08	-.05	.01	-.06
35	.56	-.54	-.15	.04	.39	-.04	.26	.10	-.04	.03	-.08
36	.39	-.38	-.10	.20	.29	-.01	.22	.03	-.13	.03	.20
37	.43	-.29	-.19	.09	.26	-.07	.13	.09	-.12	.43	-.03
38	.29	-.25	-.10	.18	.24	-.14	.28	.05	-.02	.17	.00
39	.29	-.26	-.05	.26	.35	-.03	.13	.00	-.06	.09	-.03
40	.42	.15	.15	-.08	-.10	.55	-.09	-.07	.20	.03	.01
41	.37	-.35	-.08	.34	.23	.04	.16	.07	-.10	-.05	.17
42	.62	-.69	-.16	.09	.26	-.11	.13	.05	-.00	.06	-.04
43	.37	-.24	-.34	.03	.23	.33	.03	.10	.17	-.00	-.04
44	.58	-.27	-.64	.05	.23	-.03	.14	.10	-.03	.04	-.03
45	.15	-.23	-.14	.23	.11	.01	.07	.04	-.06	.03	-.06
46	.25	-.04	-.44	.05	.14	.01	.10	.04	-.01	.09	-.07
47	.64	-.61	-.29	.11	.26	-.21	.20	.06	-.03	.10	.00
48	.60	-.67	-.18	.12	.17	-.19	.14	.02	.03	.09	-.05
49	.70	-.30	-.09	.76	.13	-.06	.03	.04	-.06	.03	-.02
50	.48	-.43	-.30	.31	.26	-.06	.10	.10	-.06	.05	.08
51	.24	.18	.14	-.04	-.35	.16	-.03	-.02	-.09	.17	.03
52	.34	.14	.15	-.03	-.37	.25	.08	-.04	.01	-.29	-.04
53	.58	-.62	-.13	.23	.23	-.01	-.00	.01	-.12	-.01	.25

Table 2. *Continued*

		Related factors loading matrix (denormalized)									
Communality		I	II	III	IV	V	VI	VII	VIII	IX	X
54	.60	-.58	-.20	.14	.30	-.22	.16	.05	-.02	.14	.13
55	.19	-.12	-.15	.04	.36	-.01	.07	-.05	-.04	.08	.08
56	.05	.01	-.02	.01	.17	.04	-.02	.05	-.05	.08	.02
Percent removed (total variance) by each of 10 factors											
		16.43	7.66	5.45	3.73	3.49	2.81	2.36	1.88	1.86	1.13
Percent total variance removed by 10 factors											
	46.80										

Table 3. Factor loading results for the 55 items for both between and within classroom

	Between							
	I	II	III	IV	V	VI	VII	VIII
1				.85				
2			+.67					
3								
5	.77							
7						-.74		
8							.49	
9							(.45)	.52
11	.87							
12	.80							
14		(56)						
15		(64)						
16					.68			
17					.73			
18			.67					
19			-.79					
20		.86						
21		(45)						
22		(70)						
23						.69		
24		(56)						
25		(40)						
26						.66		
28	.84							
30			.63					
31			.68					
33							.60	
34							.46	
35		.84						
37				.84				
39								
40					-.64			
41	.80							
42		.80						
44			+.73					
47		(67)						

Within								
I	II	III	IV	V	VI	VII	VIII	IX
	(-48)	(.42)		.46 .50		(.32)		.40
.70		.79 (.55)			.42 .45			
.75	-.67 .63						61 55	
.62 .64 .60 .62						-.60		
.59 (42) (40)	(-.68) (-.71)	(.32)				-.60		
(54)					(.20) (.27)			.43
.69 .61	-.64	(.34)		.55			(-.20)	

Table 3. *Continued*

	Between							
	I	II	III	IV	V	VI	VII	VIII
48		(67)						
49	.88							
51								-.65
52								
53	.72	(42)						
54		(66)						
55								

Within								
I	II	III	IV	V	VI	VII	VIII	IX
.67 .62		.76 (.23)	-.35 -.37					
.58			.36					

Factor I from the between analysis is identified by the following items:

5 Instructor concerned and helpful;

11 Instructor sympathetic;

12 Instructor permissive and flexible;

28 Instructor encourages criticism;

41 Instructor open to other viewpoints;

49 (Instructor sympathetic - an editorial mistake)

53 Instructor seems to enjoy teaching.

This factor describes an instructor who communicates to students a sympathetic and accepting attitude. The content of the items loading on this factor is reminiscent of the consideration factor derived from the factor analysis of employees' responses to questionnaire items describing their supervisor (Fleishman, 1979).

These same items load on Factor III from the within analysis, but somewhat lower. Thus, it appears that the between factor is <u>not</u> uniquely determined by the actual behavior of the instructor, but depends somewhat on individual differences among students in how the instructor is perceived. The "twin" items load highly on the factors from both analyses, but the items dealing with flexibility and openness load substantially less for the within than the between analysis. As a result, it appears that the "openness and flexibility" aspect of this factor is primarily attributable to the instructor.

Factor II from the between analysis is identified by the following items:

20 Well prepared

35 Uses class time well

42 Organized and presented subject matter well.

Classrooms have different students in them; these data seem to be telling us to stop asking students if they like their instructors, if their instructors are fun or maintain an atmosphere of good feeling. Instead, one should ask the blander questions related to the instructors' preparation and organization. It is obvious from the smaller, but still substantial, loadings of the items which cannot affect on this between factor, one cannot get an instructor's score independent of his ability to get students to like him (and maybe we don't want to do that, anyway), but researchers can control this aspect of the ratings to some extent by not asking the students to report their feelings.

This factor, again, is reminiscent of the Initiating Structure factor in the industrial employees' responses to their supervisor (Fleishman, 1979).

Factor III from the between analysis is identified by the following items:

2 Continue offering the course

18 Course increased appreciation for subject

19 Interested in course

30 Derived satisfaction from course

31 Good learning experience

44 Course increased knowledge and competence.

This factor is practically the same as Factor II from the within analysis and, thereby, appears to depend primarily on student perceptions rather than characteristics of the instructors.

Factor IV from the between analysis is identified by items 1 and 37, both of which deal with the textbook. These same two items occur in the within analysis (Factor XI), but the loadings are substantially lower. From this, we may infer that textbooks vary in appropriateness from course to course although there seems to be some disagreement among students in particular courses on the appropriateness of the textbook.

Factor V from the between analysis is identified by three items as follows:

16 Amount of work required is appropriate

17 Amount of material covered is reasonable

40 Amount of material covered is not too much.

The same factor occurs from the within analysis (Factor VIII) and the loadings are slightly lower. This indicates that the workload is largely a matter of student perceptions and depends little on the instructor.

Factor VI is also identified by three items:

7 Exam questions were reasonable

23 Exams reflect important aspects

26 Exams were appropriate.

The same factor occurs from the within analysis with moderately lower loadings, suggesting that there are important differences among instructors in the quality of the examinations, but students disagree somewhat on the quality of a given instructor's exams.

Factor VII is identified by three items:

8 Students are prepared for the course

33 Course is difficult enough to be stimulating

34 Temperature, lighting, etc., were comforatble.

The loadings on this factor are relatively low (<0.6), but the factor seems to appear for both the between and within analysis. Item 9, which also loads on this factor (students attend regularly), suggests, perhaps, that interested students may be less aware of minor environmentally induced discomforts. The factor is stronger for the between analysis.

Factor VIII is identified by two items

9 Students attended regularly and did assignments

51 Lectures were not repetitive of textbook.

This factor does not occur from the within analysis and suggests that instructors vary in the extent to which they derive their lectures from the textbook. The factor suggests that "too much" of this results in poor attendance.

Two other factors occur for the within, but they account for small proportions of variance and the items that load on these load as high, or higher, on other factors. As a result, reliable estimates for these factors are not available from these data. These factors are presented in Table 3.

The Analysis of Correlation and Variation of Student Rating Factors

The factors were scored by adding the transformed responses to the items indicated. These eight factor scores were analyzed further through correlations and analysis of variance.

The correlation results are presented in Table 4. Above the diagonal are the correlations derived from variability among courses. Scores for each course are computed by averaging over the students in each course and then basing the correlations on the 89 classrooms. The within correlations based on 1863 degrees of freedom reflect variability within a course where course differences are partitioned out. The within correlations are reported below the diagonal.

For each of these eight factors, an analysis of variance was done, recognizing as sources of variability, Courses (C), Sections within Courses (S/C), Sex of Student (X) and Course by Sex Interaction (CX). The computation of these analyses were also based on means rather than scores of individuals. That is, for each section, a mean for males and a mean for females were computed and these two "scores" were treated as a repeated measure of a section. That is, each section has a male score and a female score. Sections with fewer than 5 males or 5 females were not included in these analyses. Fifty-nine sections out of the 89 sections met this criterion. The results are presented in Table 5.

In order to glean from these statistical summaries the information they contain, one must integrate the two sets of results. The results are to some extent redundant and to some extent complementary. For example,

Table 4. Between classrooms (above diagonal) and within classrooms (below diagonal) correlations of the factor scores and student characteristics (df=87 for between; df=1,863 for within; variables 1-8 are student evaluations)

Within \ Between	1 Consideration	2 Instructor efficiency	3 Course effectiveness	4 Text	5 Approp. work load	6 Exams	7 Not stressed	8 Attendance	9 Sex (M=1, F=2)	10 Year	11 GPA (self-report)	12 Expected grade	13 Required (=2, Elective = 1)
1. Consideration	X[a]	55	64	39	49	53	18	37	52	08	-14	40	-31
2. Instr. Efficiency	58	X	59	45	60	56	20	36	36	-04	-02	26	-28
3. Crs. Effectiveness	45	49	X	50	48	43	43	43	58	12	-14	26	-39
4. Text	30	37	40	X	31	40	06	-04	48	02	-31	28	-33
5. Approp. wk. ld.	41	37	47	30	X	62	18	15	33	-16	-06	50	-28
6. Exams	41	41	41	32	48	X	10	27	25	-09	-07	47	-26
7. Not stressed	28	30	34	26	28	27	X	35	-09	13	20	-08	-03
8. Attendance	30	39	33	16	24	24	23	X	28	-04	06	10	-17
9. Sex (M=1, F=2)	04	06	03	07	06	04	01	05	X	-15	-22	34	-48
10. Year	02	-02	-03	-05	-04	04	01	00	-03	X	06	-11	13
11. GPA (self-report)	00	-05	06	03	05	09	06	00	19	07	X	01	16
12. Expected grade	11	05	26	13	21	27	08	02	05	00	47	X	-28
13. Required (=2, elective = 1)	01	02	01	03	-01	-02	-03	06	05	-10	-05	02	X

[a]Decimals omitted.

Table 5. F values, significance levels and error mean-squares for the analysis of variance of each of the eight factor scores

Source	df	F							
		1[a]	2	3	4	5	6	7	8
Course (C)	33	.84	1.01	3.22***	3.13***	1.03	1.64	2.14	2.10
Section (S)/C	25	9.71***	4.74***	2.27	4.70***	3.99**	4.55**	1.80	4.35**
Sex (X)	1	1.45	5.42	3.19	10.51*	4.76	5.38	1.31	3.17
CX	33	1.22	.88	1.80	1.01	1.30	1.26	.36	1.95
Error Mean Square	25	(1.824)	(.585)	(1.653)	(.168)	(.616)	(.47)	(.357)	(.171)

[a]1 = Consideration; 2 = Instructor efficiency; 3 = Course effectiveness; 4 = Text; 5 = Appropriate work load; 6 = Exams; 7 = Not stressed; 8 = Attendance.

*.01.

**.001.

***.0001.

Table 4 reveals sex differences ($p \simeq 0.05$) indicating that females tend to rate their instructors higher than do males. This is also revealed from the within-group correlations in that several of the factor scores correlate 0.05 to 0.07 with sex, which is also significant at about the 0.05 level. These correlations are trivially small, which is not apparent from the analyses of variance. On the other hand, the between correlations of the factor scores with sex are substantial, indicating that sections, or, more likely, courses, containing mostly females result in higher evaluations than courses composed of primarily males. In addition, one should note that courses where females predominate are graded higher ($r=0.34$) and are generally elective rather than required ($r=0.48$) and courses that are rated higher tend to be elective rather than required ($r=-0.28$). Further, both the student ratings and required vs. elective course are substantially related to the evaluation factors. From all of this, it seems reasonable to infer that, when males and females take the same course, they give similar evaluations, but typically female courses which tend to be elective are graded liberally and students in such courses give high evaluations. This result seems to depend little on the sex of the student judging from the low correlations from the within analysis. Rather, it seems to depend on the nature of the courses that are attractive to females (e.g., art, child development, as contrasted with chemical or mechanical engineering and computer science). Thus, it is hypothesized that, if one could entice the typical male into traditionally female courses and the typical female into traditionally male courses, their evaluations would depend little on their sex. Of course, this experiment has not been

done, and neither the males enrolled in child development courses, nor the females in engineering courses, are typical, so there is no support for this hypothesis from this observational study.

From Table 5, one observes that section differences are large, but course differences are small for Factor 1 (factor of Consideration), whereas the opposite is true for Factor 3, Effectiveness. The correlation matrix confounds these sources of variance so that one cannot directly infer whether the relationship is due to instructor (section) or course. However, section differences are usually larger than course differences, so it would seem likely that instructor variability underlies some of these relationships.

The high correlations among the eight factors from both the between and within analyses reflect the large amount of individual differences in student perceptions of instructors. That is, the two correlation matrices have the same expectations and they are, indeed, similar.

From the within group analysis, the grade a student anticipates is correlated with their evaluations, particularly for Factor 3, Effectiveness ($r = 0.26$) and Factor 6, Exams ($r = 0.27$). This suggests that within a course, the lower graded students attribute their failure to an effective instructor who gave bad tests. In contrast to this, in those courses in which students-in-general are lowly graded, the students indicate the amount of work required is too much ($r = 0.20$), as well as that the exams are poor ($r = 0.47$), but the relationship with effectiveness is relatively low ($r = 0.26$). Given the analysis of variance results, indicate that course effectiveness variability is due to the course rather than the

section, whereas amount of work is due to the section rather than the course, one might infer that instructors that demand more from their students also grade them lower! It would take more data than present research has here to test this hypothesis. The present research has only twenty-five degrees of freedom to explore covariances among variables using sections within courses and that is not worth doing.

Other than those previously mentioned, the largest difference in correlation for the between and within matrices is for expected grade and GPA. The expected grade for the between courses, $r = 0.01$ and the within courses $r = 0.47$. The latter correlation is what one should expect, but the low correlation between courses makes sense if one considers that some instructors typically grade on a curve and do not recognize the ability level of the classroom. This inference is supported by the data in that the F-ration (df = 88,1864) is 3.73 for expected grade and 1.97 for GPA. That is, relative to the variability within a classroom, classrooms differ more in the average grade given students than the average quality of the students in the class, as evidenced by their GPA. This result, along with the ones previously mentioned, does (merely) suggest that their grading practices are inconsistent and irrational and individual instructors may enhance their evaluations by grading leniently and requiring little work on the part of students.

The correlation of GPA with the evaluations tends to be low for both the within and between analysis. Between classrooms, one notes that high GPA students tend to be more critical of the textbooks ($r = -0.31$), and there is a small tendency for them to be more critical in general.

Table 6. Within correlations below the diagonal, F-ratios in the diagonal (dfs = 88,1864) and the between covariance matrix on the same scale as the within correlation matrix

	1[a]	2	3	4	5	6	7	8	9	10	11	12	13
1[a]	11.74	5.59	6.32	4.92	4.35	4.78	1.05	3.29	4.83	.98	-.27	2.63	-3.52
2	.58	8.89	5.01	4.93	4.60	4.37	1.03	2.76	2.88	-.45	-.09	1.52	-2.70
3	.45	.49	8.20	5.30	3.52	3.25	2.18	3.15	4.49	1.19	-.55	1.46	-3.69
4	.30	.37	.40	13.58	2.94	3.84	.37	-.34	4.81	.24	-1.60	2.00	-3.97
5	.41	.37	.47	.30	6.58	4.14	.80	.98	2.32	-1.37	.23	2.46	-2.30
6	.41	.41	.41	.32	.48	6.83	.47	1.82	1.78	-.79	-.27	2.37	-2.26
7	.28	.30	.34	.26	.28	.27	3.06	1.56	-.45	.80	.50	-.26	- .18
8	.30	.39	.33	.16	.24	.24	.23	6.65	1.95	-.34	.20	.49	-1.42
9	.04	.06	.03	.07	.06	.04	.01	.05	7.30	-1.43	-.82	1.79	-4.32
10	.02	-.02	-.03	-.05	-.04	.04	.01	.00	-.03	11.85	.27	-.71	1.43
11	.00	-.05	.06	.03	.05	.09	.06	.00	.19	.07	1.96	.04	.74
12	.11	.05	.26	.13	.21	.27	.08	.02	.05	.00	.47	3.73	-2.37
13	.01	.02	.01	.03	-.01	-.02	-.03	.06	.05	-.10	-.05	.02	10.71

[a]1 = Consideration; 2 = Efficiency; 3 = Effectiveness; 4 = Textbook; 5 = Appropriate work load; 6 = Exams; 7 = Not stressed; 8 = Attendance; 9 = Sex; 10 = Year; 11 = GPA; 12 = Expected grade; 13 = Required.

However, there seems to be a statistically significant, but small, trend in the opposite direction within classrooms, particularly in regard to Examinations ($r = 0.09$). Significantly (in both senses $r = -0.05$, $p = 0.02$) high GPA students within classrooms tend to regard their instructors as less efficient. Since the other evaluations tend to correlate positively with GPA, this negative trend merits further exploration using a better indicator of ability than the GPA used in this study.

Focusing on students who take required versus elective courses, one observes little difference among the two groups of students within courses, as evidenced by the consistently low correlations of this dichotomous variable with the remaining ones. However, between courses, one observes that students taking required courses tend to get lower grades ($r = -0.28$), tend to be male, and evaluate the course lower than students who take elective courses. This result indicates it would be unfair to compare evaluations between instructors where their classrooms differ in composition on this variable.

In summary, it appears that student characteristics play a very minor role in determining how students evaluate the instructor within any given classroom with the rather obvious exception that higher graded students evaluate their instructor higher. However, when instructors teaching different courses are analyzed, the characteristics of the classroom are strongly related to the evaluations. The strongest relationships occur in regard to sex, so that one may infer that courses that are selected by women have better instructors or women who select traditionally feminine courses give higher evaluations. These data are not adequate to address

the causation question, but they do clearly demonstrate that there are serious problems associated with comparing evaluations from instructors teaching different kinds of student, or instructors who use different grading standards.

SUMMARY

This chapter summarizes the study and offers conclusions, suggestions for implementation, and recommendations for further research in four sections. The first section of this summary concerns purposes and procedures. Major findings and suggestions for further research are summarized in the second section, and the third discusses overall findings and suggestions for further research. Finally, recommendations for appropriate evaluation by student ratings are given for the benefit of administrators and faculty members.

Purposes and Procedures

The major objectives of this research were to determine how variances in student ratings relate to student characteristics, such as sex, expected grade, etc., and to study other variables which could affect student responses in rating different courses.

This research was intended to analyze the student ratings to achieve that objective by employing the between-class data analysis and the within-class data analysis. Doing so provides the results which are based on instructor level (through the B-analysis) and student level (through the W-analysis). Therefore, the reports of this study are the reports which are based on those two levels.

The student evaluation form for faculty was developed for use as an instrument in this study. Samples, consisting of 88 classrooms which were from the multiple section courses taught by different instructors in a variety of disciplines at Iowa State University, were collected

during spring quarter in 1981. The total number of students in the samples was 2,107.

The between-class data and within-class data were analyzed by the MANOVA procedure (SAS). The factor extraction procedure (PRINIT) and the rotation procedure (VARIMAX) also were implemented by SAS.

Results

Interpretation of the results of the between-class and within-class data analyses showed that student ratings related to many variables. All variables were found to associate with each other and to affect student ratings, as discussed below.

1) Sex of student shows an effect on student ratings in the B-analysis, but not in the W-analysis. In the W-analysis, correlation is small, though significant at the 0.05 level, and appears in many factors, although it is not apparent in the analysis of variance. Therefore, one could conclude from the W-analysis that male and female students within a class tended to rate the instructor the same, but the B-analysis showed a tendency by female students to rate the instructor higher than do males. It must be noted, however, that this correlation appeared in courses which were attended predominantly by females, or in courses which were attractive to females, and which were elective, rather than required, courses. Thus, it appears that sex of the student alone is not a major contributor to higher ratings of instructors, but that other variables also have some effect. In other words, sex of student appears to be just one of several variables which affect student ratings, and it is possible that this variable has even less effect considering that the results from the

W-analysis showed small correlation. It would seem that students taking a course as an elective course would favor that particular course and this favorability is probably more of a contributing factor to the ratings the students give than whether or not they are male or female students. However, further research is needed to prove this, and could be done using a sample from typically male-oriented courses with some female majors, and samples from traditionally female-oriented courses with some male majors, to determine whether male and female students will rate the instructors differently or not.

2) It was found that the elective courses got higher ratings than the required courses.

3) The different results of student ratings appeared greater within sections, compared to between courses. This means that instructors affect the results of student ratings, rather than courses. In other words, the results of student ratings depend upon the instructors' teaching effectiveness regardless of the courses they are teaching. This researcher believes that this result is quite acceptable. It is true that some courses are easier than others. However, this does not mean that the instructors of the harder courses will always get lower ratings from students than instructors of easier courses. It may be only a tendency. In general, students are aware of the difficulty of the subject matter being taught, and if the instructor demonstrated excellence in teaching effectiveness, then the results of the student ratings would not be low. On the other hand, if an instructor teaching a less difficult course did not demonstrate effective teaching, the results of the student ratings would be

low. Therefore, this researcher accepts, to some degree, the results which show that the instructor affects the students' ratings rather than the difficulty of the particular course. This does not mean that course effect is nonexistent, but that it has less weight than instructor effect. Further research is necessary to obtain support to this finding regarding teaching effectiveness versus course taught. Samples should be obtained from different courses (hard and easy) which have the same groups of students showing common characteristics, such as major, GPA, reason for taking the course (required vs. elective), and results of student ratings compared and studied, whether instructors or courses affect student ratings more. In general, it is not suggested that the results of student ratings of different courses be compared, because there are other variables reflected in the outcome besides the instructor's teaching effectiveness. The results of student ratings should be compared only when those courses have the same groups of students, such as described. Then, the results might be compared.

4) Low grade students attribute their failure to poor exams and ineffective instructors.

5) Both the B-analysis and the W-analysis showed that high GPA students did not rate the instructor higher in the effectiveness factor and may even evaluate the instructor as being less effective. It showed that high GPA students were more critical of texts and instructors. However, when considering the overall student ratings, it was found that the student ratings were positively correlated with GPA. This suggests that it might be desirable to obtain the actual grade of the students for the

purposes of further study. In this study, the student GPA and expected grade was reported by each student, since the subjects were anonymous. It would seem that the actual grade received could be a better indicator than the GPA. Therefore, it is suggested that future studies obtain the actual grade of the student.

6) The expected grade and student ratings were found to have a lower correlation in the B-analysis and higher in the W-analysis. The low correlation in the B-analysis is more varied. The students might expect to receive a higher grade from easier courses than from harder courses. Also, it revealed that the student taking a required course intended to get a lower grade compared to an elective course. The higher correlation of expected grade and student rating found in the W-analysis is acceptable. It could be due to the fact that students who expected higher grades in class were the students who had an effective instructor, or an instructor was an easier grader. It is necessary to know the actual grade for further research, but, in this study, the researcher could obtain only the grade expected by the student. This researcher recommends that, for further study, an attempt be made to obtain the actual grade received, even though a study done by Maxey and Ormsby (1971) for the American College Testing Program showed a high correlation between the expected grade and the final grade actually received.

Conclusions

The overall findings are listed below.

It was found that, in all eight factors of student ratings, the individual differences in students were reflected in student perception of instructors. The results reveal that both B-analysis and W-analysis help to obtain the fact of that perception. In other words, looking at both analyses can help to obtain the actual evaluation of instructors by students.

This study revealed that the student characteristics played a minor role in student ratings of their instructors within a course, but when more classes were studied, the student characteristics played a stronger role in the student ratings, especially in the courses that contained more female students, feminine types of courses, and elective courses.

It was also shown that all variables in this study were associated with each other and affected higher student ratings of some courses. This study could not show which variables are the actual causes for higher ratings of a class. This could be considered a limitation of this study, which was intended only to investigate the results of student ratings as affected by many variables when looking at both the between- and within-course data. If the experimental research was performed, it could reveal what is or are the actual cause(s) or effect(s) of higher or lower student ratings.

Suggestions for Further Research

Most of the suggestions for future research appeared in the summary of findings.

One additional suggestion that this researcher would like to make is that future research should obtain more samples. It seemed that the multiple courses which were the samples of this study may not be necessary. This should provide more 400-level courses, which the present study lacked. It is suggested that administration of the questionnaire be done in a uniform manner.

Suggestions for Administrators and Faculty Members

Suggestions for the student rating form are as follows. First, the form should contain a number of items which the faculty member could use for the purpose of improvement of his teaching. The faculty members are the primary beneficiaries in that they receive feedback from the student ratings. It was found that some forms do not contain enough items to provide adequate information to the faculty. The results of student ratings from these forms are of little value when the administrators also ignore the results or use them in an inappropriate manner. Therefore, any student evaluation form should be for use by the faculty. It might be desirable to have not only a universal form for an entire university, but also a particular form for a particular discipline. The different disciplines have different teaching-learning processes. Therefore, the use of a student evaluation form designed for use in a particular discipline would be more appropriate and beneficial than a universal one.

Administrators should use student evaluation of faculty as a part of their decision-making process, especially in the case of an extreme decision regarding a faculty promotion. The record of student ratings should be kept consistently. The consistency of ratings (either favorable or unfavorable) over a period of several years time should be taken into account for promotion and/or improvement of instructors and courses. However, the administrators have to be aware that the results from student ratings are only a part of the total system for teaching evaluation which could be used as reference regarding faculty.

After conducting this study, this researcher continues to believe that student evaluation is a useful tool for evaluation of teaching. At present, student ratings are also useful in program planning, curriculum development, etc. Design and development of student evaluation forms for particular purposes provid a greater benefit from the use of student ratings. Some application studies of student ratings have emerged in recent literature, which shows the expansion in use of student ratings for other purposes, as well as for teaching evaluation. Even though this study found that many variables influence student ratings, this does not mean that student ratings' results are useless or unreliable. It is true that student ratings are affected by some variables, but they do still contain major information of teaching evaluation for faculty and administration. The faculty members and administrators must compromise the results of student ratings with all of the information about the students, which are the variables that affect the student ratings. This information consists of student sex, year standing, reason for taking course (required or

elective), GPA, expected grade, and major. Knowledge of this information would help both faculty and administrators to make compromises and to have greater understanding of the results of the student ratings.

REFERENCES

Aleamoni, Lawrance, M. 1981. Student rating of instruction. *In* Jason Millman (ed.) Handbook of teacher evaluation. Sage Publications, Beverly Hills, California.

Aleamoni, L. M. and N. H. Graham. 1974. The relationship between CEA ratings and instructor's rank, class size, and course level. Journal of Educational Measurement 11:189-201.

Aleamoni, Lawrance M. and P. Z. Hexner. 1980. A review of the research on student evaluation and a report on the effect of different sets of instruction on student course and instructor evaluation. Instructional Science 9:67-89.

Aleamoni, Lawrance M. and R. E. Spencer. 1973. The Illinois course evaluation questionnaire: A description of its development and a report of some of its results. Educational and Psychological Measurement 33: 669-684.

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

Astin, A. W. and C. B. T. Lee. 1967. Current practices in the evaluation and training of college teachers. *In* C. B. T. Lee (ed.) Improving college teaching. Amerian Council on Education, Washington, D.C.

Barnoski, R. P. and A. L. Sockloff. 1976. A validation study of the faculty and course evaluation (FACE) instrument. Education and Psychological Measurement 56:391-400.

Bauswell, R. B. and C. R. Bauswell. 1979. Student ratings and various instructional variables from a within-instructor perspective. Research in Higher Education 11:167-177.

Bauswell, R. B., E. Schwartz and A. Purohit. 1975. An examination of the conditions under which various student ratings parameters replicate across time. Journal of Educational Measurement 12:273-280.

Bejar, I. I. 1975. A survey of selected administration practice supporting in student evaluation of instruction program. Research in Higher Education 3:77-86.

Bejar, I. I. and K. O. Doyle. 1974. Generalizability of factor structures underlying student ratings of instruction. Research report presented at the meeting of the American Educational Research Association. Chicago, Illinois. ERIC ED 693 945.

Bendig, A. W. 1952. A preliminary study of the effect of academic level, sex and course variables on student rating of psychology instructors. Journal of Psychology 34:21-26.

Bendig, A. W. 1953. An inverted factor analysis study of student-rated introductory psychology instructors. Journal of Experimental Education 21:333-336.

Bendig, A. W. 1954. A factor analysis of student ratings of psychology instructors on the Purdue Scale. Journal of Educational Psychology 45: 385-393.

Blum, M. M. 1936. An investigation of the relation existing between students' grades and their ratings of an instructor's ability to teach. Journal of Educational Psychology 27:217-221.

Brandenburge, D. C., J. A. Slinde and E. G. Batista. 1977. Student ratings of instruction: Validity and normative interpretations. Research in Higher Education 7:67-78.

Brown, D. L. 1976. Faculty ratings and student grades: A university wide multiple regression analysis. Journal of Educational Psychology 68:573-578.

Bryan, R. C. 1968. Student rating of teachers. Improving College and University Teaching 16:200-202.

Cattell, R. B. 1978. The scientific use of factor analysis in behavioral and life sciences. Plenum Press, New York.

Centra, John A. 1973. Self-ratings of college teachers: A comparison with student ratings. Journal of Educational Measurement 10:287-295.

Centra, John A. 1980. Determining faculty effectiveness. Jossey-Bass Publisher, San Francisco, California.

Centra, John A. and R. L. Lin. 1976. Student points of view in ratings of college instruction. Educational and Psychological Measurement 36: 693-703.

Coffman, W. E. 1953. A preliminary attempt to identify the factors in student-teacher evaluation. Journal of Psychology 36:417-422.

Coffman, W. E. 1954. Determining students' concepts of effective teaching from their ratings of instructors. Journal of Educational Psychology 45:277-285.

Cohen, J. and L. G. Humphrey. 1960. Memorandum to faculty. Unpublished manuscript. Department of Psychology, University of Illinois, Champaign-Urbana, Ill.

Cohen, S. A. and W. G. Berger. 1970. Dimensions of students' ratings of college instructors underlying subsequent achievement on course examination. *In* Proceedings of the 38th Annual Convention of the American Psychological Association, 1970 5:605-606.

Cosgrove, D. S. 1959. Diagnostic rating of teacher performance. Journal of Educational Psychology 50:200-204.

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

Crannel, C. W. 1953. A preliminary attempt to identify the factors in student-teacher evaluation. Journal of Psychology 36:417-422.

Crichton, L. I. and K. O. Doyle. 1975. Reliability of ratings. Measurement Service Center, University of Minnesota, Minneapolis, Minn.

Cronbach, L. J. and N. Webb. 1975. Between-class and within-class effects in a reported aptitude x treatment interaction: Re-analysis of a study by G. L. Anderson. Journal of Educational Psychology 67:717-721.

Cronbach, L., G. Gleager, H. Nanda and N. Rajaratnan. 1972. The dependability of behavioral measurement: Theory of generalizability for scores and profiles. Wiley, New York.

Deaux, K. and Taynor, T. 1973. Evaluation of male and female ability: Bias works two ways. Psychological Report 32:261-262.

Desphande, A. S., S. C. Webb and E. Marles. 1970. Student perception of engineering instructor behaviors and their relationships to the evaluation of instructors and courses. American Education Research Journal 7:289-305.

DeWolf, V. A. 1972. Student ratings of instruction in post secondary institutions: A comprehensive annotated bibliography of research reported since 1968 (Vol. 1). University of Washington, Bureau of Teaching, Seattle, Washington.

Dickinson, T. L. and Leroy Wolins. 1974. Least square analysis of repeated measures and other designs. Multivariate Behavioral Research 9: 353-371.

Doyle, K. O., Jr. 1975. Student evaluation of instruction. Lexington Books, D. C. Heath and Company, Lexington, Mass.

Doyle, K. O., Jr. and L. I. Crichton. 1978. Student, peer and self evaluations of college instructors. Journal of Educational Psychology 70:815-826.

Doyle, K. O., Jr. and S. E. Whitely. 1974. Student ratings as criteria for effective teaching. American Educational Researcher Journal 11: 259-274.

Elliott, D. H. 1950. Characteristics and relationships of various criteria of college and university teaching. Purdue University Studies in Higher Education 70:5-61.

Elmore, P. B. and LaPointe, K. A. 1975. Effect of teacher sex, student sex, and teacher warmth on the evaluation of college instructors. Journal of Educational Psychology 67:368-374.

Elmore, P. B. and LaPointe, K. A. 1978. Effect of teacher, student, and class characteristics on the evaluation of college instructors. Journal of Educational Psychology 70:187-192.

Elrod, M. M. and S. J. Crase. 1980. Sex difference in self-esteem and parental behavior. Psychological Report 46:719-727.

Feldman, K. A. 1976. Grades and college students' evaluations of their courses and teacher. Research in Higher Education 4:69-111.

Feldman, K. A. 1977. Consistency and variability among college students in rating their teachers and courses: A review and analysis. Research in Higher Education 6:223-274.

Feldman, K. A. 1978. Course characteristics and college students' ratings of their teachers: What we know and what we don't. Research in Higher Education 9:199-242.

Finkbeiner, C. T., J. S. Lathrop and J. M. Schwerger. 1973. Course and instructor evaluation: Some dimensions of a questionnaire. Journal of Educational Psychology 64:159-163.

Fleishman, E. 1979. The description of supervisor behavior. Journal of Applied Psychology 37:1-6.

French-Lazovik, G. 1974. Predictability of students' evaluation of college teachers from component rating. Journal of Educational Psychology 66:373-385.

Frey, P. W. 1973. Student ratings of teaching: Validity of several rating factors. Science 182:83-85.

Frey, P. W., D. W. Leonard and W. W. Beatty. 1975. Student ratings of instruction: Validation research. American Educational Research Journal 12:435-444.

Gange, N. L. 1974. Student ratings of college teaching: Their justification and proper use. *In* N. S. Glasma and B. R. Killait (eds.) Second UCSB Conference on Effective Teaching. University of California, Santa Barbara, California.

Garverick, O. M. and H. D. Carter. 1962. Instructor ratings and expected grades, California Journal of Educational Research 13:218-221.

Gillmore, G. M. 1973. Estimates of reliability coefficients for items and subscales of the Illinois Course Evaluation Questionnaire. Research Report N. 341. Office of Instructional Research, Measurement and Research Division, University of Illinois, Urbana, Ill.

Gillmore, G. M. and D. C. Brandenburge. 1974. Would the proportion of students taking a class as a requirement affect the student ratings of the course? Research Report No. 347. Office of Instructional Research, Measurement and Research Division, University of Illinois, Champaign-Urbana, Ill.

Gillmore, G. M., M. T. Kane and R. N. Naccarato. 1976. The generalizability of student instructional ratings: General theory and application to the University of Washington Instructional Assessment System (EAC Reports), Educational Assessment Center, University of Washington, Seattle.

Gillmore, G. M., M. T. Kane and R. W. Naccarato. 1977. The teacher and the course as unit of analysis in the generalizability of student ratings of instruction. Project 77-9:506, Educational Assessment Center. University of Washington, Seattle.

Gillmore, G. M., M. T. Kane and R. W. Naccarato. 1978. The generalizability of student ratings of instruction: Estimation of the teacher and course components. Journal of Educational Measurement 15:1-3.

Glasman, S. N., B. R. Killait and W. Gmelch. 1974. Evaluation of instructors in higher education: Concepts, research, and development. University of California at Santa Barbara.

Gordon, G. E., C. M. Bridges, B. William and J. E. McLean. 1973. Student evaluation of college teaching behavior instrument: A factor analysis. Journal of Higher Education 44:596-604.

Graham, Margaret H. P. 1973. Relationship of selected characteristics of student evaluation of college instruction. Unpublished dissertation research. University of Illinois, Champaign-Urbana, Illinois.

Granzin, K. L. and J. J. Painter. 1973. A new explanation for students' course evaluation tendencies. American Educational Research Journal 10:115-124.

Grush, J. E. and F. Costin. 1975. The student as consumer of the teaching process. American Educational Research Journal 12:55-66.

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

Halstead, J. S. 1970. A model for research on ratings of courses and instructors. Proceedings of 78th Annual Convention of the American Psychological Association 5:695-626.

Harari, O. and S. Zedeck. 1973. Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology 58:261-265.

Hendricks, R. L. 1974. The effects of response format and internal versus external criteria measures on the evaluation of importance weighted models of job satisfaction. Unpublished Ph.D. dissertation. Iowa State University, Ames, Iowa.

Hildebrand, M. H., R. C. Wilson and E. R. Dienst. 1971. Evaluating university teaching. Research and Development in Higher Education. University of California, Berkeley, Calif.

Hogan, T. P. 1973. Similarity of student ratings across instructors, courses, and time. Research in Higher Education 1:149-154.

Holmes, D. S. 1971. The relationship between expected grades and students' evaluations of their instructors. Educational and Psychological Measurement 31:951-957.

Holmes, D. S. 1972. Effects of grades and disconfirmed grade expectations on students' evaluation of their instructors. Journal of Educational Psychology 63:130-133.

Howard, G. S. and S. E. Maxwell. 1980. Correlation between student satisfaction and grades: A case of mistaken causation. Journal of Educational Psychology 72:810-820.

Isaacson, R. L., W. J. McKeachie and J. E. Milholl. 1963. Correlation of teacher personality variables and student ratings. Journal of Educational Psychology 54:110-117.

Isaacson, R. L., W. J. McKeachie and J. E. Milholl. 1964. The dimensions of student evaluations of teaching. Journal of Educational Psychology 155:344-351.

Johnson, J. 1976. An analysis of student evaluation of course and instructor on an expression of the students' values and educational attitudes. Unpublished Ph.D. dissertation. North Carolina State University.

Kaiser, H. F. 1958. The Varimax criteria for analytic rotation in factor analysis. Psychometrics 25:187-200.

Kane, M. T. and R. Brennan. 1977. The generalizability of class means. Review of Educational Research 47:267-292.

Kane, M. T., G. M. Gillmore and T. J. Crook. 1974. The application of generalizability theory to course evaluation questionnaires. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Chicago, Ill.

Kane, M. T., G. M. Gillmore and T. J. Crook. 1976. Student evaluations of teaching: The generalizability of class means. Journal of Educational Measurement 13:171-184.

Kennedy, W. R. 1975. Grades expected and grades received either relationship to students' evaluation of faculty performance. Journal of Educational Psychology 67:109-115.

Kent, L. 1966. Student evaluation of teaching. The Educational Record 47:376-400.

Knapper, C. K., L. G. George and others. 1977. If teaching is important: The evaluation of instruction in higher education. Canadian Association of University Teachers, Clark, Irwin and Company, Ltd., Canada.

Kulik, J. A. and W. J. McKeachie. 1975. The evaluation of teachers in higher education. *In* F. N. Kerlinger (ed.). Review of Research in Education. Peacock Publications, Ithaca, N.Y.

Leventhal, L., P. C. Abrami, and R. P. Perry. 1976. Do teacher rating forms reveal as much about students as about teachers? Journal of Educational Psychology 68:441-445.

Linn, R. L., J. A. Centra and L. Tucker. 1975. Between, within and total group factor analysis of student ratings of instruction. Multivariate Behavioral Research 10:242-288.

Lolli, Anthony, Jr. 1977. A uni-wide multivalidative of a student rating scale for teacher evaluation. Unpublished Ph.D. dissertation. University of Connecticut.

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

Marsh, H. W. 1978. Students' evaluations of instructional effectiveness: Relationship to student, course, and instructor characteristics. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, March 1978 ERIC ED 155 217.

Marsh, H. W. and J. W. Overall. 1979. Long-term stability of students' evaluations: A note on Feldman's consistency and variability among college students in rating their teachers and courses. Research in Higher Education 10:139-147.

Marsh, H. W., J. W. Overall and C. S. Thomas. 1976. The relationships between students' evaluation of instruction and expected grade. University of California, Los Angeles, Calif. ERIC ED 126 140.

Marsh, H. W., J. W. Overall and P. S. Kersler. 1979. Validity of student evaluations on instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. Journal of Educational Psychology 71:149-160.

Mathew, J. B. 1975. Assessment of students evaluation procedures as assures of teaching effectiveness. *In* Improving University Teaching, International Conference, Heidelberg, Germany, May 9-11.

Maxey, J. E. and V. J. Ormsby. 1971. The accuracy of self-report information collected on the ACT Test Battery: High school grades and items of nonacademic achievement. The American College Testing Program, Research and Development Division. ACT Research Report No. 45. Iowa City, Iowa.

McKeachie, W. J. 1975a. Correlates of students variables. *In* A. L. Sockloff (ed.) Proceedings Faculty Effectiveness as Evaluated by Students. Temple University, Philadelphia, Penn. ERIC ED 101 469.

McKeachie, W. J. 1975b. Assessing teaching effectiveness: Comments and summary. *In* Improving University Teaching, International Conference, Heidelberg, Germany, May 9-11.

McKeachie, W. J., Y. Kin and W. Mann. 1971. Student ratings of teacher effectiveness: Validity studies. American Educational Research Journal 8:435-445.

McNemar, A. 1962. Psychological Statistics. 3rd ed. Wiley Publishing Company, New York.

Millman, J. (ed.). 1981. Handbook of teacher evaluation. Sage Publications, Beverly Hills, Calif.

Morsh, J. E., G. G. Burgess and P. N. Smith. 1956. Student achievement as measurement of instructor effectiveness. Journal of Educational Psychology 47:79-88.

Nadeau, G. 1974. Student evaluation of instruction: The rating questionnaire. *In* Naomi E. S. Griffiths (ed.) The Evaluation of instruction of University Teachers, Monograph Series. Clark, Irwin and Company, Canada.

Nichols, M. G. 1969. A study of influences of selected variables involved in student evaluation of teacher effectiveness. Unpublished dissertation, University of South Dakota.

Pohlmann, J. T. 1975a. A multivariate analysis of selected class characteristics and student ratings of instruction. Multibehavioral Research 10:81-91.

Pohlmann, J. T. 1975b. A description of teaching effectiveness as measured by student ratings. Journal of Educational Measurement 12:49-53.

Powel, R. W. 1977. Grade and student evaluation of instruction. Research in Higher Education 7:193-205.

Rader, N. F. 1968. College student ratings of instructors. The Journal of Experimental Education 37(2):76-81.

Remmers, H. H. 1960. Manual of instruction for the Purdue Rating Scale for Instructors (rev. ed.) Purdue University, West Lafayette, Ind.

Russell, H. E. and A. W. Bendig. 1953. Investigation of the relations of student ratings of psychology instructors to their course achievement when academic aptitude is controlled. Educational and Psychological Measurement 13:626-635.

Sirotnix, K. A. 1980. Psychometric implications of the unit of analysis problem. Journal of Educational Measurement 17:245-282.

Smith, P. L. 1979a. The generalizability of student ratings of courses: Asking the right questions. Journal of Educational Measurement 16:77-87.

Smith, P. L. 1979b. The stability of teacher performance in the same course over time. Research in Higher Education 11:153-165.

Sockloff, A. L. 1975. Behavior of the product moment correlation coefficient when two heterogeneous subgroups are pooled. Educational and Psychological Measurement 35:267-276.

Solomon, D. 1966. Teacher behavior dimensions, course characteristics, and student evaluations of teachers. American Educational Research Journal 3:35-47.

Solomon, D., L. Rosenberg and W. G. Bezdek. 1964. Teacher behavior and student learning. Journal of Educational Psychology 55:23-30.

Spencer, R. E. and L. M. Aleamoni. 1970. A student course evaluation questionnaire. Journal of Educational Measurement 7:209-210.

Stumpf, S. A. 1979. Assessing academic program and department effectiveness using student evaluation data. Research in Higher Education 11: 353-363.

Stumpf, S. A. and R. D. Freedman. 1979. Expected grade deviation with student ratings of instruction: Individual versus class effects. Journal of Educational Psychology 71:293-302.

Tetenbaum, T. 1977. The factor invariance of student ratings of instruction under three sets of directions. Research in Higher Education 6: 11-23.

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

Weigel, R. G., R. O. Oetting and L. O. Tasto. 1971. Differences in course grades and student ratings of teacher performance. School and Society 99:60-62.

Whitely, S. E. and K. O. Doyle. 1976. Implicit theories in student ratings. American Educational Research Journal 13:241-253.

Whitely, S. E. and K. O. Doyle. 1978. Dimension of effective teaching: Factors or artifacts. Educational and Psychological Measurement 38: 107-117.

Whitely, S. E. and K. O. Doyle. 1979. Validity and generalizability of student ratings from between-classes and within-class data. Journal of Educational Psychology 41:117-124.

Wolins, L. In press. Research mistakes in the social and behavioral sciences. Iowa State University Press, Ames, Iowa.

Wolins, L. and T. L. Dickinson. 1973. Transformation to improve reliability and for validity for effective scales. Education and Psychological Measurement 33:711-713.

Wotruba, T. R. and P. L. Wright. 1975. How to develop a teacher-rating instrument: A research approach. Journal of Higher Educational Psychology 46:653-663.

Yonge, G. D. and J. M. Sassenrath. 1968. Student personality correlates of teacher ratings. Journal of Educational Psychology 59:44-52.

ACKNOWLEDGEMENTS

There are many individuals to whom I wish to express my gratitude:

To sources of financial support -

- P.E.O. members who made possible the P.E.O. International Peace Scholarship, and to the members who made my stay in the United States warm with their friendship, love and caring.
- Mr. Suppachi, who assisted me with many things through his great love and kindness, and made my wish of coming to the United States to continue my education become possible.
- Mrs. Suchada, my sister, who first gave support to me to study.

To Advisors and Committee Members -

- My present advisor, Dr. Larry Ebbers, and my former advisor, Dr. Ray Bryan, for their guidance and encouragement throughout my graduate work. My degree would not have been possible without their help. In addition, my appreciation is given to my committee members, Dr. Trevor G. Howe, Dr. Ted J. Solomon, and Dr. Leonard G. Smith.

To the Staff of the Statistics Department -

- Dr. Leroy Wolins, who deserves special thanks and recognition for the many times he generously gave assistance. I sincerely thank him. Without his ideas, encouragement, and direction during the statistical analysis, this study could not have been done. An additional note of appreciation is given to Dr. Bud J. Meador and his assistant, Mr. Shih Tai-Tao for the programming assistance they gave with the data analysis.

APPENDIX A.

THE QUESTIONNAIRE FORM

1	50	99
Strongly Disagree	Neutral	Strongly Agree

If you STRONGLY DISAGREE with the statement given, place a 1 in the blank. If you disagree but believe it isn't a "STRONG" disagreement, choose any number between 1 and 50, registering the extent to which you disagree with the statement.

If you STRONGLY AGREE, place a 99 in the blank; however, if you agree but do not believe that you "STRONGLY AGREE," choose a number between 50 and 99 to show the extent to which you agree with the statement.

If you neither agree nor disagree, that is, you are completely neutral about a statement, use the number 50.

**You are encouraged to use any of the numbers from 1 to 99 in your ratings.

If the statement does not apply in your situation, place 50 in the answer blank.

Respond Here

____ 1. The textbook(s) used in this course were appropriate in content and difficulty.

____ 2. This course should continue to be offered.

____ 3. I was often confused.

____ 4. The instructor's objective for the course has been made clear.

____ 5. The instructor seemed genuinely concerned with students' progress and was actually helpful.

____ 6. The instructor stimulates intellectual curiosity.

____ 7. Exam questions were unreasonably detailed (picky).

____ 8. The students in this course generally attended the lectures and did the assignments.

____ 10. The grading was fair.

____ 11. The instructor displays a sympathetic attitude to students.

____ 12. The instructor is permissive and flexible.

(continued)

Respond Here

____ 13. There was considerable agreement between the announced course objective and what was actually taught.

____ 14. I would like to take another course from this instructor.

____ 15. The instructor for this course was better than most other instructors I have had.

____ 16. The amount of work required is appropriate for the credit received.

____ 17. The amount of material covered in the course is reasonable.

____ 18. The course increased my appreciation for the subject.

____ 19. I have little or no interest in the content of this course.

____ 20. The instructor was well-prepared for each class.

____ 21. The instructor maintains an atmosphere of good feeling in the class.

____ 22. The instructor made good use of examples and illustrations.

____ 23. Examinations reflect the important aspects of the course.

____ 24. The comments I have heard from other students about the instructor were favorable.

____ 25. The instructor made the course entertaining and fun.

____ 26. Overall, the exams were appropriate.

____ 27. The method or methods by which subject matter of the course (recitation, lecture, laboratory, etc.) was presented was appropriate.

____ 28. The instructor encourages constructive criticism.

____ 29. The instructor seemed enthusiastic about the subject matter.

____ 30. I have derived much satisfaction from taking this course.

____ 31. This course was a good learning experience.

(continued)

Respond Here

____ 32. The instructor gave assignments that were useful for learning subject matter.

____ 33. The instructor has made the course sufficiently difficult to be stimulating.

____ 34. The physical environment (temperature, lighting, etc.) in the classroom were reasonable comfortable.

____ 35. The instructor uses class time well.

____ 36. Class discussion was welcome.

____ 37. The textbook and classroom instruction complemented each other well.

____ 38. The instructor told the students how they would be evaluated in this course.

____ 39. The instructor was available to the students at times other than during lectures or labs.

____ 40. The instructor attempted to cover too much material.

____ 41. The instructor was open to other viewpoints.

____ 42. The instructor organized and presented subject matter well.

____ 43. Compared to other courses, I have put in much effort in this.

____ 44. This course has increased my knowledge and competence in this area.

____ 45. The instructor made available supplementary material (instructional aids, references, etc.) to students who were interested or needed them.

____ 46. The content of this course is relevant to my major field.

____ 47. In my opinion the instructor has accomplished (is accomplishing) his or her objectives for the course.

____ 48. The instructor spoke understandably, or explained the subject clearly.

____ 49. The instructor displayed sympathetic attitude to students.

(continued)

Respond Here

____ 50. The instructor demonstrated the importance and significance of the subject.

____ 51. The lectures were too repetitive to what was in the textbook(s).

____ 52. The lectures seemed unrelated to what was in the textbook(s).

____ 53. The instructor seemed to enjoy teaching.

____ 54. The instructor achieved (or is achieving) the specified objectives of the course.

____ 55. I have given thoughtful consideration to the questions on this form.

Please indicate the response to the following questions by circling the appropriate response.

56. What is your sex? Male Female

57. What year are you in? Freshman Sophomore Junior Senior

58. What is your Quality Point Average (or Grade Point Average)? _____

59. What is your grade in this course at the present time?

A A- B+ B B- C+ C C- D+ D D- F

60. What is your major? _____

61. This course is: an elective a required course

61. Have you had a course from this instructor previously? Yes No

APPENDIX B.

FACTORS FROM THE BETWEEN ANALYSIS

Factor I - Consideration

5 The instructor seemed genuinely concerned with students' progress and was actually helpful.

11 The instructor displays a sympathetic attitude to students.

12 The instructor is permissive and flexible.

28 The instructor encourages constructive criticism.

41 The instructor was open to other viewpoints.

49 The instructor displayed sympathetic attitude to student.

53 The instructor seemed to enjoy teaching.

Factor II - Instructor Efficiency

20 The instructor was well-prepared for each class.

35 The instructor uses class time well.

42 The instructor organized and presented subject matter well.

Factor III - Course Effectiveness

1 This course should continue to be offered.

18 The course increased my appreciation for the subject.

19 I have little or no interest in the content of this course.

30 I have derived much satisfaction from taking this course.

31 This course was a good learning experience.

44 This course has increased my knowledge and competence in this area.

Factor IV - Text

1 The textbook(s) used in this course were appropriate in content and difficulty.

32 The textbook and classroom instruction complemented each other.

Factor V - Appropriate Amount of Work

16 The amount of work required is appropriate for the credit received.

14 The amount of material covered in the course is reasonable.

40 The instructor attempted to cover too much material.

Factor VI - Difficulty of exams

7 Exam questions were unreasonably detailed (picky).

23 Examinations reflect the important aspects of the course.

26 Overall, the exams were appropriate.

Factor VII Stress of Students

8 The students in this course appear to have adequate prior preparation to learn the subject matter.

33 The instructor has made the course sufficiently difficult to be stimulating.

34 The physical environment (temperature, light, etc.) in the classroom were reasonably comfortable.

Factor VIII - Attendance

9 The students in this course generally attended the lectures and did the assignments.

51 The lectures were too repetitive to what was in the textbook(s).