In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.
Abstract

Despite problems with methodology and interpretation of results, student evaluations of instruction provide useful data for administrative decisions. This study specified and tested an Hierarchical Linear Model that can be used to ameliorate some of these problems. The study examined the student ratings of 3,689 university courses taught by 260 instructors that were collected over an 11-year period in an education faculty. A longitudinal hierarchical linear model was used to investigate whether individual instructors' scores were stable over time and in a variety of contexts. The model was also used to examine the effects of exogenous course and instructor variables on the scores.

Results showed that the effects of the course-level variables class size, course level, percentage of females in the class and percentage of students taking the course as an elective were significant. Together these variables accounted for approximately 20% of the within-instructor variance and 17% of the between-instructor variance. The effects of the instructor-level variables rank, sex and number of course taught were not significant, although the sex of the instructor was substantively related to both the average score and the course level.

Scores increased significantly over time, but with less reliability. The years of experience of the instructor who taught the course was also significant. The analysis also illustrated the utility of this model for score adjustment and recommendations were made with respect to the use of the scores for summative evaluation.
# Table of Contents

Abstract ii  
Table of Contents iii  
List of Tables iv  
List of Figures v  
Acknowledgment vi  

Chapter One: Introduction  
1.1 Background of the Problem 2  
1.2 Statement of the Problem 5  
1.3 Research Questions 10  

Chapter Two: Review of Literature 11  
2.1 Scores Used for Summative Evaluation 11  
2.2 Potential Sources of Bias 16  
    Student Characteristics 17  
    Course Characteristics 22  
    Instructor Characteristics 25  
    Implications for the Use of Student Ratings 28  
2.3 Reliability 29  
    Longitudinal Designs 37  

Chapter Three: Research Methodology 43  
3.1 Subjects and Data Collection 43  
3.2 Measures 44  
    Dependent Variables 44  
    Independent Variables 45  
3.3 Data Analysis 47  
    Hierarchical Linear Models 47  
    Models in the Analysis 50  

Chapter Four: Results 53  
4.1 Descriptive Statistics 53  
4.2 Null Models 55  
4.3 Course and Instructor Models 62  
4.4 Longitudinal Models 72  

Chapter Five: Summary 81  

References 88  

Appendix 1: Annual Teaching Evaluation Questionnaire 98
List of Tables

Table 1: Means and Standard Deviations of the Composite Score and Global Item by Course Level, Rank and Sex of Instructor 53

Table 2: Descriptive Statistics for Course and Instructor Variables 55

Table 3: HLM Results for the Null Models of the Composite Score and Global Item 56

Table 4: HLM Analysis of the Effects of Course-Level Variables on Course Means Within Instructors 63

Table 5: HLM Analysis of the Effects of Instructor-Level Variables on Course-Level Means Within Instructors 69

Table 6: HLM Analysis of the Effects of Course-Level Variables on Course Means Within Instructors: Time Model 74

Table 7: HLM Analysis of the Effects of Course-Level Variables on Course Means Within Instructors: Instructor Experience Model 79
List of Figures

Figure 1: Reliability of SET Scores 59

Figure 2: Predicted SET Score by Class Size 65

Figure 3: Predicted SET Score over Time 73

Figure 4: Predicted SET Score by Years of Experience 77
Acknowledgment

The three members of my committee provided unique contributions to this thesis, and I would like to thank them for their help and guidance.

Dr. Kjell Rubenson provided suggestions with respect to the sociological and policy implications of this work.

Dr. Frank Echols was grace personified throughout the completion of this work and generously offered his time for a detailed examination of the thesis, valuable feedback and continued support for my academic career.

Special thanks to Dr. Doug Willms, my advisor, for his guidance and feedback in all aspects of this thesis. Dr. Willms was a patient and tolerant teacher and his offer of such a large data set to this masters student was extraordinary, and will not be forgotten. In addition, his statistical expertise lead to a more sophisticated analysis than would have originally been undertaken for a project at this level.

Finally, I must thank my family for their moral support. Their contribution enabled me to maintain proper perspective, a sense of humor and the energy to create a healthy balance while completing this thesis. I would also like to thank Hank and Clara. Because of them I’m still standing.
Chapter One: Introduction

Student ratings of instruction have been used with increased frequency and acceptance in North American universities over the past 20 years. Both political factors and extensive research have contributed to this increase in use (Murray, Rushton, & Paunonen, 1990). Administrators obtain valuable information about the quality of instruction from the ratings, and also benefit politically from the provision to students of “a legitimate and satisfying role in evaluating their educational experiences and those who provide them” (Gillmore, 1983, p.557).

Research has supported the use of students ratings of instruction through reports that the ratings are reasonably stable across items, raters and time periods; are minimally affected by extraneous factors such as class size, severity of grading, and instructor characteristics; are consistent with ratings given by alumni, colleagues, and trained classroom observers; and are significantly correlated with measures of student achievement such as standardized tests (Marsh, 1984; Murray, 1980). The research is primarily North American, but in European countries public pressure to demonstrate quality in higher education has lead to interest in the use of student ratings of instruction and related research (Husbands and Fosh, 1993).

Marsh (1987), a prolific academic and strong advocate for the use of student ratings, reviewed many years of research and concluded that student evaluations of teaching were: (a) multidimensional; (b) reliable and stable; (c) primarily a function of the instructor who taught the course rather than of the course that was taught; (d) relatively valid against a variety of potential biases; and (e) seen to be useful by faculty for feedback about their teaching, by
students for use in course selection, by administrators for use in personnel decisions, and by researchers for data on effective teaching (p.1). Marsh and others (e.g., McKeachie, 1979, 1986; Murray, Rushton & Paunonen, 1990) have supported the use of student ratings as a measure of teaching effectiveness upon which to base administrative decisions about faculty tenure, retention, salary, and promotion. Critics have suggested, however, that there is, or should be, considerable debate about the legitimacy of their use, and the methods used for their inclusion in faculty evaluation (Husbands and Fosh, 1993; Murray, 1987). This controversy over the use of student ratings informs this study, which will present an Hierarchical Linear Model that evaluators can use to mediate some of the issues pertaining to the use of student ratings of instruction for personnel decisions.

1.1 Background of the Problem

Student evaluations of teaching are usually structured questionnaires that require students to evaluate instructors on multiple-choice or rating scale items regarding clarity of instruction, organization of material, evaluation procedures, and facilitation of student participation. Students spend considerable time with an instructor and are therefore in a good position to judge the both the quality of the course and the quality of the instruction. Students also provide a useful service to the university administration, in that ratings can be acquired in an efficient and relatively inexpensive manner (Gillmore, 1983). The results of these evaluations are generally used for two purposes: formative evaluation and summative evaluation (Murray, 1987).
In summative evaluation the results of student ratings provide data for administrative decisions about faculty tenure, promotion, retention, salaries, and merit pay. Student ratings should provide a valid and reliable measure of teaching effectiveness if they are to be used in making personnel decisions. In formative evaluation the results of student ratings provide instructors with diagnostic feedback to improve their teaching. As feedback, results obtained from the ratings should provide a valid diagnosis of the strengths and weaknesses of the instruction and should be accompanied with suggestions for improvement (McKeachie, 1979). Because the purpose of each type of evaluation is distinct, the two processes should be kept separate both conceptually and procedurally (Murray, 1987).

Cashin and Downey (1992) observed that formative evaluation and summative evaluation are often confounded conceptually in the research literature. The purpose of summative evaluation is to make an accurate assessment about how effectively the instructor taught, while controlling for exogenous variables other than the effectiveness of the instructor. In this case, the job of the evaluator is neither to ascertain why the instructor was successful or unsuccessful nor to make suggestions for improvement. This is the role of the formative evaluator who often operates within the context of faculty development programs focused on improving teaching and researching the teaching process. In order to avoid confusion over the interpretation of the results of research, the purpose of the evaluation procedure in question must be clearly defined.

A further distinction between the two types of evaluation arises out of faculty responses to the use of ratings. Feedback used for improving teaching may appear to be neutral and is perhaps even welcome by some faculty. However, the use of ratings for
personnel decisions may be perceived as threatening, particularly in situations where decisions based on these data seriously affect the lives and careers of faculty, and where the decisions are final or difficult to change (Miller, 1982). Discomfort with the evaluation process often gives rise to objections and considerable controversy over the inclusion of data from student ratings in faculty portfolios compiled for the review process.

In a study by Murray et al. (1982), faculty members were surveyed as to their views about the use of student ratings of instruction. The results indicated that although 78% of the faculty rated feedback as a very important goal for the use of student evaluations, only 42% rated their use for administrative decisions as highly important. Thus faculty, who are often involved in evaluating their colleagues, may not be motivated to place emphasis on the use of student ratings. Faculty members also hold personal views from the perspectives of both the evaluator and the evaluated that may influence how research about these ratings is conducted. Faculty researchers have studied the effects of factors unrelated to instruction or particular courses in order to confirm the validity of student ratings and have used the rating instruments to clarify what makes good teaching, but they have also undertaken this type of research in order to ‘debunk’ student ratings (Gillmore, 1983, p.568).

Faculty members often object to: 1) the undue influence given to quantitative data which appears alongside narrative, subjective evaluations; (2) the error of measurement of student rating scales, which may be large enough to render the rating score meaningless for making decisions about individuals; (3) the lack of accompanying information about the meaning, interpretation, and limitations of the scores; and (4) the fact that promotion and tenure decisions are often made by administrators and faculty committees who do not
understand the standard criteria for measurement instruments, do not know how to interpret the results, or do not realize the limitations of the results (Miller, 1982, p.88). The use of student evaluations persists since:

We live in social contexts in which all of us are judged by some standard by someone. The teaching-learning process is difficult to capture with words or on a rating scale, yet objective treatment of the process is much fairer and more defensible than mystical pronouncements about it. It would be a genuine mistake, however, for those advocating more objective and systematic faculty evaluation systems to take lightly this and other objections to faculty evaluation, and especially to take lightly the objection to student ratings of classroom teaching. (Miller, 1986, p.166)

This type of evaluation is problematic, but it would be premature to abandon the use of student ratings and lose a valuable source of information. In light of faculty objections, the challenge for evaluators is to produce a rating system that appears equitable and takes into account exogenous variables which may affect the scores, and to provide administrators with results that are both interpretable and of practical significance.

1.2 Statement of the Problem

Many authors close their studies with suggestions for the fair use of student ratings of instruction for making personnel decisions. The research issues that inform these suggestions include 1) the choice of an appropriate measure or score; 2) potential sources of bias;
3) reliability of the scores; and 4) the development of a system for score adjustment. However, in spite of numerous suggestions and considerable effort invested in research related to the first three issues, with regard to adjustment of the scores “little has been done to explore procedures for handling them when employing student evaluations for making serious, practical decisions about faculty merit” (Shingles, 1977, p.459). Perhaps little has been done because of the complexity of the task and the inevitably politically charged controversy that surrounds this evaluation process. However, the development of an appropriate rating system is important because:

Accurate information as to how well faculty perform their teaching role is vital to administrators who must make decisions on tenure, promotion and salaries; to faculty whose careers and lives are directly affected by these decisions and who may profit from constructive criticism; and to students who ultimately experience the consequences of any such evaluation in the classroom. (Shingles, 1977, p.459)

Husbands & Fosh (1993) observed that for formative purposes, some unreliability in the scores may be unimportant. By contrast, for important decisions about faculty such as decisions about careers and monetary rewards, even a small amount of unreliability becomes a serious matter. For faculty at the extremes of the distribution, who have either very good or very poor evaluations, some inaccuracy may be acceptable. However, the issue of the accuracy of the scores “arises in using evaluation results mechanistically to make fine discriminations between individuals in the modal sections of a distribution” (p. 110). Accuracy of the scores is also important in situations where the results are compared to some
criterion or norm determined by the evaluators. The question then becomes: if there are extraneous factors that affect the reliability of the scores, how should they be accounted for in the evaluation process?

Considerable research has examined student ratings for possible sources of bias which might affect their accuracy. Faculty have been concerned that extraneous factors over which they have no control may affect the classroom environment and students’ perceptions, and therefore negatively affect evaluations. Although this area of research has been prolific, researchers continue to call for systematic research into correlates, and continue to collect and code data on a variety of background variables in order to control for the effects of exogenous variables (Husbands & Fosh, 1993). Some researchers have suggested that if these factors influence the scores, they must be taken into account before the ratings are used to make important decisions (Marsh, Overall, and Kesler, 1979). A more precise point of view is that the use of unadjusted scores is questionable if student ratings of instruction are influenced by and correlated with such factors (Shingles, 1977). To date the degree to which certain factors affect the scores has yet to be examined in conjunction with an attempt to control for these factors and adjust the scores to produce standardized norms that provide objective assessments of teaching effectiveness based on student evaluations of instruction.

Results of student evaluations can also be applied to the assignment of teaching duties, wherein instructors are assigned to teach the type of course in which they receive more favorable evaluations. This practice implies that the assigned courses are the courses in which the individual excels as an instructor. The possibility that some courses are rated higher than others by virtue of their content or curriculum is not taken into account. In
addition, the potential to create bias in the scores exists if all faculty can not be accommodated by this system. Senior faculty will have an advantage over novice faculty members, since new instructors do not have ratings on which to base the assignment of their teaching duties.

The situation becomes more complex when researchers suggest that administrators should take into account additional contextual information such as the differences between types of courses, the teaching methods used for a particular course, and the circumstances in which the courses were taught (McKeachie, 1986). To ameliorate the effects of course differences and make a fair assessment of the effectiveness of the instructor, McKeachie (1979) suggested that ratings should be obtained from more than one class and more than one type of course for each instructor. Erdle, Murray, and Rushton (1985) concurred, noting that a faculty member who was denied tenure or promotion on the basis of an inadequate sample of their teaching, could possibly make the case that had they been assigned other types of courses, their evaluations would have been more favorable.

Cranton & Smith (1986) noted that student ratings data are often collected in order to rank instructors or courses based on class means, and to develop norms that can be used for comparisons between instructors. However, external factors may not have the same influence on scores from different contexts and the policies about the use of the scores may differ across departments, faculties and universities. Thus, it may be difficult to weight factors external to the instruction equally in situations where comparisons between instructors are to be made (Miller, 1982). Cranton & Smith (1986) suggested that evaluators take into account different course characteristics and consider giving weight to individuals who teach in
contexts which usually get lower ratings. In their view, university-wide or even faculty-wide norms should not be used, and instructors teaching in similar contexts should be compared only after the development of situational norms based on class size, course level and department.

Abrami (1985) also had these concerns, and suggested that a model of effective instruction based on interaction terms, as in aptitude and treatment interaction research (ATI), should be promoted. This would require the development of complex rating systems which take into account the ecological conditions of instruction such as setting, course and student characteristics and the outcomes of instruction. The inclusion of such contextual data further complicates the task of devising a system which is accessible and interpretable by administrators. The quantity of information required for this type of evaluation would become unmanageable without the use of a quantitative tool such as the surveys. As Miller (1986) noted, “quantification is desirable in order to process large amounts of data, but these data always need to be buttressed with systematic and reasoned judgment” (p. 166).

Systematic judgment is particularly important when interpreting the scores.

The purpose of this study is to specify and test an Hierarchical Linear Model that is used to address these issues and to provide administrators with a tool to analyze the results of student evaluations of teaching for use in personnel decisions. The study examines student ratings of faculty who have been evaluated over an eleven-year period with the same evaluation form. It employs a longitudinal hierarchical linear model to investigate whether individual instructors’ scores are stable over time and in a variety of contexts, and the effects of exogenous course and instructor variables on the scores. This research also illustrates the
utility of this type of model for score adjustment when results of the evaluation are to be used as summative feedback for administrative decisions.

1.3 Research Questions

The study will address three main questions:

1. How many courses should be used to provide a reliable assessment of an individual instructor's score?

2. a. Do exogenous course and instructor characteristics significantly affect student ratings?

2. b. If these exogenous characteristics are of statistical significance, is it of practical significance to adjust the scores for personnel decisions?

3. Does a longitudinal model provide additional information which contributes to theory and is of practical significance when considering adjustment of the scores?

The second chapter reviews the literature related to reliability, biases and the use of scores from student evaluations of instruction for administrative decisions. Chapter three outlines the sample, data collection and analysis used in the study. Chapter four presents the results of the analyses and Chapter five discusses the findings arising from this investigation.
Chapter Two: Review of Literature

This chapter reviews literature related to three issues that are relevant to the use of student evaluations of teaching for summative evaluation. The first issue concerns the selection of an appropriate outcome measure. Procedures and measures under debate include the choice of a global question or a composite score, the use of factor scores, and the use of a weighted mean. The second issue concerns the identification of factors that may bias the scores, and the procedures used to account for these factors. The third issue concerns the reliability of the scores, and the extension of this research to include a longitudinal model.

2.1 Scores Used for Summative Evaluation

A recent topic of discussion in the literature is the selection of the most appropriate score or scores to be used for administrative decisions (Hativa & Raviv, 1993). Many student rating questionnaires contain items related to particular aspects of effective teaching and global items which ask students to rate the overall effectiveness of the instructor or the course. The question under debate is whether it is better to use a single score (a global item or the mean of the individual items) which represents teaching effectiveness as a single comprehensive dimension, or multiple scores which represent different dimensions of teaching. Johnson (1989) found from a survey of experts that equal numbers favored and opposed the use of a global rating as the sole measure of teaching effectiveness to be used for summative evaluations.
One problem with the use of an average of the items is that aggregation of the data results in loss of useful information and loss of meaning. McKeachie (1979) argued that while an aggregate score simplified interpretation, the use of a single score implied that teaching was unidimensional and that each item on the instrument had equal weight within the construct of effective teaching. However, he also observed that the use of many single items in an analysis increased the possibility of some results being significant by chance (Type I error) resulting in the loss of reliability. He therefore favored the use of factor analysis to consolidate the items into separate dimensions of teaching and reduce the number of items in the analysis.

Marsh (1984, 1985, 1987) also suggested that factor analysis be used to reduce the number of scores reported to several dimensions. In summaries of research based on his own instrument Marsh (1991a; 1991b) concluded that since teaching is a multifaceted process, measures of teaching effectiveness should also be multidimensional. Therefore research on his instrument, the SEEQ, is based on a nine-factor model that includes the dimensions learning/value, instructor enthusiasm, organization/clarity, group interaction, individual rapport, breadth of coverage, examinations/grading, assignments/readings and workload/difficulty. He opposed the use of a single criterion of effective teaching, speculating that each dimension or factor may correlate differently with indicators of effective teaching (such as students’ performance indicators) and may be affected differently by exogenous variables. These speculations remain unconfirmed by research (Abrami, 1989a).
Abrami (1989a; 1989b) argued in favor of the use of multiple scores for feedback to improve teaching, and the use of a single score for administrative decisions. He suggested that rating scale research had yet to provide sufficient evidence to conclude that the true dimensions of effective teaching have been identified and correctly interrelated. Because of the lack of consistency in factor analyses, insufficient theories about teaching and lack of content validity of some items across different contexts it would premature to rely on dimensions of instruction identified through factor analysis as the scores to be used for summative evaluation.

Abrami (1989a, 1989b) observed that a single score may be the most appropriate choice if the ability of administrators to weigh information from factor scores is questionable. Administrators might weigh the factor scores equally, search out only the strongest or weakest areas for a particular instructor, or place more emphasis on factors which they personally favor over others. These practices would contribute subjectivity and inconsistency to the evaluation process. One would, however, expect administrators to be able to interpret multiple scores since they routinely must weigh the value of complex data. In addition, proper information about the meaning, interpretation and limitation of the scores would reduce the possibility of this problem. Abrami favored the use of the global item, but in addition cited two other common choices for single scores: (1) a simple unweighted mean score of all rating items except the global items; and (2) a weighted mean of all items, selected items, or factors identified through a factor analysis procedure (p.626).

A global item may be easier to use than an item mean or weighted-item mean, but whether a global item is truly representative of the mean of all items is questionable. McBean
and Lennox (1987) and Hativa and Raviv (1993) both described the concern and puzzlement expressed by faculty when the global score was higher or lower than the item mean. Arguably, the global question may not of be value if it does not appear to summarize the responses of the individual questions and if faculty have no indication of how to improve the global rating if it is lower than the item mean.

In response to the concern of faculty about the apparent differences between the two scores, McBean & Lennox (1987) investigated whether a linear combination (or average) of the individual items would provide different results than the global rating. They recognized the downside of using the linear combination, in that the average included the assumption that the items were independent and additive, but also recognized the need to provide a single summary score to administrators. When the items were averaged and compared to the mean of a global question using regression analysis, the two scores were highly correlated (.86), but did not have a slope of 45 degrees. The authors concluded that since the t-test showed that there was little likelihood of the slope being unity, the aggregate score and the global item could not be used interchangeably. They recommended the use of the linear combination of items in order to avoid negative responses by faculty to the global item on the questionnaire.

Hativa and Raviv (1993) also compared aggregate scores and global items and found that all correlations of the global score with the item mean were above .95, but some correction was required in order to use the global item to replace the item mean. The authors suggested that it may be easier and more practical to use the global item if high correlation
exists between the two measures, but the process of correcting the global score reduces the
ease of its use.

One of the conceptual problems with a global item is that it asks students to
differentially weigh the components of teaching effectiveness in arriving at their response
(Marsh, 1991b). In theory they are being asked to intuitively weigh the individual items in
order to arrive at an overall score. Proponents of this type of question appear to assume that
all students arrive at a score for the global item using the same criteria. If this assumption
were untrue, the validity of these items would become an issue. This is an interesting point in
light of the previous suggestion that administrators, who are not trained in evaluation, may
not be able to appropriately consider and weigh single items or factor scores that are
presented to them directly. This caveat has not been discussed in the literature with respect to
the process of scoring such items, perhaps because students only provide the data and are not
in a position to make decisions.

Both Abrami (1989a) and Marsh (1991b) conceded that a weighted item average
might be the best compromise, but have yet to provide a method which operationalizes this
type of average. They cautioned that administrators often lack the knowledge and expertise of
instructional evaluators and must be provided with an accurate and defensible procedure for
any type of synthesis of the score. For example, Marsh suggests that scores which are not
considered to be of importance, such as workload or course difficulty, should be given a low
or zero weighting. This suggestion raises a question about the content validity of such items:
if some items on the questionnaire should be given a zero rating, should they appear on the
questionnaire at all?
Another suggestion is that guidelines for weighting the items should be dependent on the particular context in which the ratings take place, taking into account the type of course, the goals of the course, and so forth. This becomes a complex process, particularly if the resulting scores are to be used to compare faculty members or academic units. The concept of weighting the mean may be appealing as a way to ameliorate the issue of which score to select for summative evaluation, but empirical and policy-related research is required in order to demonstrate how to arrive at the appropriate weights.

The selection of the most appropriate score should be related to the practice of the department in which the evaluation takes place, the content and construction of the instrument and the reporting practices in a particular setting. Researchers will no doubt continue to favor several different summary scores derived from student ratings of instruction and continue to debate which score is the most appropriate to be used in making personnel decisions. Researchers should take into account contextual factors at the site of the evaluation and examine the descriptive results of the evaluation before decisions are made as to the most appropriate outcome measure.

2.2 Potential Sources of Bias

Researchers opposed to the use of student evaluations of teaching maintain that since the ratings are biased they do not reflect true differences in teaching effectiveness (Hofman & Kremer, 1980; Machlup, 1979; Ryan, Anderson, and Birchler, 1980). Others maintain the case for bias is overrated (Abrami & Mizener, 1983; Barke, Tollefson & Tracy, 1983; Centra, 1977; Doyle & Whitley, 1974; Sullivan and Skanes, 1974) and have concluded that student
evaluations do reflect real differences in teaching effectiveness (McKeachie, 1986; Marsh 1987; Marsh & Dunkin, 1992).

Bias may be conceptualized as "the influence of variables reflecting something other than the instructor's teaching effectiveness" (Cashin & Downey, 1992, p.565). Considerable research on student evaluations of teaching has involved the examination of factors which may bias the scores and are outside the influence of the instructor (Shapiro, 1990). However, once these variables are identified, it becomes difficult to demonstrate that variables which are highly correlated with student ratings have a causal effect on the ratings. Statistical techniques such as multiple regression have been used to address this issue.

Research into potential biases persists, since the possibility of their influence raises questions about the validity of the scores, the interpretation of the scores, and the utility of the scores for administrative decisions in higher education (Husbands and Fosh, 1993). The possibility that the influence of some factors is negligible and of no practical significance, while the influence of other factors is considerable and has serious consequences lead these researchers to state that "notwithstanding Marsh's generally positive views on the subject, it is clear that the issue of potential biases still arouses strong emotions among many who are being, or about to be, evaluated" (p.104). Potential biases can be grouped into three general categories of characteristics: 1) student; 2) course; and 3) instructor (Shapiro, 1990).

**Student Characteristics**

Student characteristics studied include sex, age, level of program, and personality traits. These variables have been consistently found to have little effect on student ratings
Class grades, which have been defined as student, instructor and course characteristics, have been correlated with student ratings. Generally higher course grades are correlated with higher student ratings, and Brown (1976) found that 8 to 9% of the variance in student ratings could be attributed to course grades. In a summary of previous correlational studies, Feldman (1976) found the relationship between ratings and grades to vary from .20 to .50. However, Cohen (1981) discovered that correlations between ratings and achievement were higher (.86) when students knew their final grade at the time of the evaluation than when they did not (.38).

Three causal connections have been suggested for the relationship between grades and ratings (Gillmore, 1983). First, a good grade acts as positive reinforcement and causes the students to like the course better and give it a higher rating. This is an example of a course-level bias. Second, if the student likes the course they tend to give it a higher rating, and are also more motivated to do well in the course. This is an example of both an instructor effect and bias in operation. The instructor motivates the student, and the student brings a bias in terms of liking the course. Third, there may be factors that are a result of the student. The student may have an inherent interest in the subject matter which causes both higher ratings and higher grades. In addition, the effects of student expectations on ratings have also been studied and results showed that students who expected a course or teacher to be good tend to rate the course higher (McKeachie, 1979). A fourth explanation argued in favor of the expertise of the instructor, suggesting that good teachers encourage student learning, which
should lead to higher grades (Howard and Maxwell, 1980). This is what we would hope to find, that the instructor effect alone influences the grades.

The correlation between course grades and student ratings was described by Marsh (1984) as a "grading leniency effect". He proposed a student characteristic model in which good students select good teachers, and therefore good grades became associated with good instruction through the initial course selection process. In addition, good instructors are teaching students with higher motivation and greater interest (p.737). However this model has yet to be tested and does not take other explanations into consideration. It could be that the students who give higher ratings are not so much motivated to learn as they are to search out the instructors who are known for giving higher grades. This model also assumes that students have a choice in the selection of particular courses and instructors. The problem with investigating such models is that a variable such as course grades cannot be manipulated in non-experimental research.

Ideally one would expect student ratings to reflect quality of instruction, and course grades to reflect the amount of student learning. Assuming students learn more when they are taught well, we would therefore expect student ratings and course grades to be positively related. But the relationship will be imperfect not only if student ratings inadequately reflect quality of instruction, but also if grades are an invalid measure of what was learned (Baird, 1987). Subjective evaluations of how much the student thought they learned may more accurately predict student ratings than letter grades. Baird compared a question about the amount the student thought they had learned to the overall instructor rating, the overall course rating, and the anticipated letter grade. A small correlation of .18 was found between
course grades and perceived learning. Correlations between course grades and the overall ratings of the professor and the course were modest (.28 and .33). However correlations between perceived learning and the overall ratings of the professor and the course were much higher (.86 and .88), the implication being that ratings are more a function of the perceived quality of learning than the expected grade in the course.

Studies which investigate the effects of student grades often have mixed results (Gillmore, 1983). The variety of study designs which use the ratings at the individual level or aggregate the data at the course level could account for some of the inconsistency, and pooling groups of classes may mask much of the variability which exists from class to class. Indeed this could be the explanation for the lack of significant findings with respect to many other student variables. One of the problems in studying student effects is that when class means are used to look at teacher effects, students from several classes are pooled, incorporating the variability due to students into the between-class differences for the instructors. Thus information about the individual students is lost (Abrami & Mizener, 1983).

Several studies have examined the personality characteristics of students and attitude similarity to the instructor. Abrami, Perry, and Leventhal (1982) studied the relationship between student personality characteristics and course ratings and found that these characteristics did not account for a significant amount of the variance in ratings. However they did find that student perceptions of the instructors' personality were related. They found the students' perception of the instructors' need to achieve to be a good predictor, but generally found that for making "gross distinctions" among instructors this should not be an important factor. They concluded that students did not appear to let their attitudes or their
attitude similarity with the instructor affect the evaluation and found that students were able
to effectively discriminate between the instruction and their perceptions of teachers’ attitudes.
A subsequent study by Abrami & Mizener (1985) reported results consistent with prior
research. Modest correlations (mean $r = .23$, $p < .05$) were found between student ratings and
perceived attitude similarity and these relationships were reduced by 56.2% when the
influence of instructors on the ratings was removed. Therefore these variables did not
contribute to substantial bias in the ratings.

Studies of personality characteristics of both students and instructors may be of
interest in terms of the quality of instruction and for matching students to the most
appropriate instructors, but measures of these types of variables are difficult to include as a
regular part of the evaluation process. First, collection of this type of data may make the
evaluation process more time consuming, and students may not wish to provide this type of
data, even if the evaluations are anonymous. Second, since there is no strong evidence to
indicate that students learn more from professors with a certain type of personality, it would
be a questionable practice to rate faculty on selected personality traits of themselves and their
students and compare results based on such assessments. This practice may suggest that there
is a “correct” personality or set of behaviors that an instructor must possess or aspire to in
order to obtain higher ratings.

Most studies of student evaluations use undergraduate students as subjects. Of interest
is a study by Shapiro (1990), in which off-campus nontraditional students were studied.
Results indicated that these students responded in a manner similar to traditional university
students. Husbands and Fosh (1993) suggested a number of other student characteristics
which could be considered in future studies. For example, determining whether the students are “mature” students, indigenous or “overseas” students, undergraduate or postgraduate students, full or part-time students or are in the early or later phase of their program might provide variables of interest. Other variables include the country of origin of the student, their competence in the language of instruction, whether they are paying their own fees or are in receipt of some type of support, their numeracy or literacy in the subject matter and the political orientation of the teacher vis-à-vis the student.

Course Characteristics

Course characteristics which have been studied include class size, difficulty of the course, time of day the class met, class grades, how and when the evaluations were administered and the purpose of the evaluation. Ratings are higher when the students know that the evaluations are to be used for tenure or promotion decisions, when the evaluation is not anonymous, or when the instructor is present during the evaluation (Marsh, 1984).

In a review of the literature, Feldman (1978) examined the effects of course level, elective or required course and the type of course, and found that ratings were higher in upper level courses and elective courses and were higher in the humanities than the sciences. The time of day the class met did not affect the ratings. Subsequent research supported the finding that both course level and whether the course was required had an effect on the scores, with the highest ratings in advanced and elective courses (Marsh, 1980). The differences between the ratings of upper and lower division classes is well-established in the literature (Cranton
and Smith, 1986; Whitten and Umble, 1980). Marsh (1984) also found that the difficulty of
the course affected the ratings, with difficult courses rated more highly.

Feldman (1984) summarized more than 50 studies on the effects of class size and
found that although one third of the studies showed no relationship with student ratings, the
majority of the studies found a small inverse relationship. He concluded that class size
explained between 1 and 8% of the variance in student ratings. Several studies reported
curvilinear relationships in which medium-sized classes were rated lower than small and
large classes, but other researchers found no significant differences between ratings given by
students in small, medium and large classes (see Aleaomi & Graham, 1974).

One study examined the relationship of class size to overall ratings and factor scores
using polynomial trend analysis (Marsh, Overall & Kesler, 1979). Four questions about class
size and student ratings informed this research: 1) Is the relationship linear or nonlinear? 2) Is
the effect uniform across different components of students' evaluations? 3) Can the
relationships be “explained away” by controlling for other variables describing the student,
the course, or the instructor? and 4) What are the practical implications of the results in the
use of students' evaluations? ( p.58). The results showed that the relationship was clearly
nonlinear with the largest and smallest classes receiving higher ratings. A fourth order
polynomial best described the relationship and in two separate studies accounted for 5.5%
and 3.5% of the variance. However the polynomial relationship was more difficult to define
in samples with very few large classes over 200, indicating that it may not be beneficial to
focus on the large extremes when there are few cases in the sample and the effect of class size
is small.
Several reasons for the relationship between the scores and class size have been suggested (Marsh, Overall & Kesler, 1979). For example, instructors may have been assigned to very large classes on the basis of previously demonstrated success teaching the course or because of a personal interest in the course. For large classes, it may have been that introductory or prerequisite courses had no enrollment limits and therefore grew in size because of the reputation of the instructor. If that were the case then the expectation of receiving good instruction would have influenced the ratings, not the class size. It may also have been that instructors in very large classes were more highly motivated or challenged because of the large size of the class.

Other researchers have confirmed the results of the effects of course characteristics (Cranton & Smith, 1986), but have found that in specific departments, or for particular levels of instruction these relationships varied. Therefore the results depended on the context in which the ratings were collected. The researchers used four overall questions about the instruction, the amount learned, the significance of the learning and the value of the course to examine the effects of being a full-time or part-time instructor, in a day or an evening course, the particular campus (two sites), the level of instruction and class size. They found no differences between full and part-time instructors or campuses, but found a small effect size for being in a day course which was not consistent over all four departments (.10 to .18). They also found consistent trends of higher ratings in higher levels of courses and in smaller classes, but when the variables were examined by department there was considerable variation in the degree of these trends. These results suggested that the contextual effects of
departments and faculties should be taken into consideration when examining the effects of these variables on the scores.

Husband and Fosh (1993) suggested additional characteristics which could be examined at the course level such as: service course in students' own department; class outside their own department; method of assessment; teaching environment; gender ratio of the teaching context; full semester course; and the length of class sessions.

*Instructor Characteristics*

Instructor characteristics such as sex, rank and personality characteristics have not shown consistent relationships with student ratings (McKeachie, 1979) and researchers have concluded that student evaluations are not affected by the instructor's sex, publishing record, personality traits and faculty rank (McKeachie 1979; McKeachie, 1986; Marsh, 1984). Other instructor characteristics studied include age of the instructor and teaching experience.

Dukes and Victoria (1989) examined the effects of sex and rank of the professor and found they had no significant effect on student ratings. They suggested two factors that may account for the lack of gender bias reported in the research. First, students may be sufficiently competent to rate instructors by focusing on the quality of teaching rather than on the characteristics of the instructor. Second, society ascribes high status to university professors (as seen in scores on measurements of occupational prestige), and this ascribed status may transcend any potential gender bias in the ratings. However, it may be that studies which examine the effect of sex of the instructor in particular contexts may yield different results,
particularly in faculties where there are traditionally few male or female instructors or where there are extremes in the ratios of male and female students.

Empirical studies relating student ratings and research productivity have consistently found a small positive relationship (Feldman, 1987; Rushton et al., 1983). However little attention has been paid to research productivity and its relationship to teaching effectiveness in different types of courses (Murray, Rushton and Paunonen, 1990). Findings were obtained that suggested that instructor ambitiousness, endurance, compulsiveness, and intelligence correlated significantly with student ratings in graduate seminars and undergraduate methodology courses. These results combined with the previous finding that these personality traits correlated significantly with faculty publication rates and citation counts (Rushton et al., 1983) implied that research productivity was positively related to teaching effectiveness, but only for certain research-oriented courses. It is questionable, however, as to whether the publication rate stands alone as a factor which influences scores or merely serves as a proxy for the personality traits mentioned above.

In a meta-analysis of studies relating to seniority, age and experience of the instructor, Feldman (1983) reported that in two-thirds of the studies reviewed there was no significant relationship between faculty rank and student ratings. About one-half of the studies found no significant relationship between the ratings and the age and teaching experience of the instructor. When relationships were found, higher ranked faculty received higher ratings, but there was a consistent negative relationship with teaching experience. Some evidence for a curvilinear relationship between ratings and experience was found, with an increase in ratings
over the first few years of an instructors’ career, but this was quite weak over larger aggregations of faculty.

Feldman (1986) found no relationship between student ratings and instructors’ self-reported personality traits but found a relatively strong relationship between student ratings and personality traits as assessed by student or colleague perceptions. Erdel, Murray and Rushton (1985) concurred and viewed these results as evidence that ratings converge with theoretical variables about teaching. Contrary to the claim that ratings were influenced by the personality of the instructor, the results also suggest that instructor personality is reflected in specific classroom teaching behaviors which in turn are validly rated by students.

The personality characteristics of the instructor may have some bearing on the ratings, and may be a factor in effective instruction. However, the characteristics are not exogenous variables and the use of this type of data in conjunction with student ratings and personnel decisions would be rather controversial. Abrami & Mizener (1985) noted that it would be unusual to consider instructor effects as biasing the results, since the ratings are supposed to reflect differences in instructional effectiveness. The effects described as instructor effects include the effects of teaching unique to the instructor, and also the variance shared with other effects that influence the scores. Therefore, findings from studies such as Erdle and Murray (1986) on teaching behaviors, Murray, Rushton & Paunonen (1990) on personality type, and Erdle, Murray & Rushton (1985) on classroom behaviors are important with regard to faculty development, but would not be taken into consideration when comparing faculty or adjusting scores. The instructor variables of interest for score adjustment are those factors which contribute to the variance in the scores but are outside of the control of the instructor.
Implications for Research and Use of the Scores

The use of multivariate analysis techniques has been suggested as a way to ameliorate the inconsistent findings between exogenous factors and student ratings. These inconsistencies may occur because researchers do not usually study a number of characteristics simultaneously (Whitely & Doyle, 1979). Both Shingles (1977) and Centra (1993) argued that factors must be examined together since the cumulative effects of a number of factors may result in the unfair use of scores for decision making. Many factors may have a small effect on their own, but when their influence is taken together with other variables there may be a substantial effect on the scores. Inconsistent and nonsignificant findings in the literature may prompt the researcher to eliminate these variables from their study. In reality, for their particular context, the cumulative effects of many factors may result in significant relationships and lead to substantial adjustment of the scores.

Researchers advocate taking such factors into account when developing comparative norms and creating weighting systems for adjusting the scores (Cranton & Smith, 1986), and suggest that there is a need to adopt a norm referenced model of instructional evaluation (Abrami & Mizener, 1985). This allows teaching quality to be judged relative to the performance of an appropriate peer or norm group such as the faculty or department to which the instructor belongs. Criterion referencing requires the a priori establishment of instructional standards independent of the relative performance of a teacher’s peers. Separate norm groups may be formed according to class size, class level and student motivation to control for extraneous effects on scores.
Abrami & Mizener (1985) caution that the creation of norm groups may provide the "illusion of precision" by controlling for extraneous factors which contribute to the variance of the scores. However, these norms may be based only on those factors which are easily measured and may not include more intangible variables. Others have suggested that over the long run, when ratings for an individual instructor are averaged over many courses, the estimate obtained will provide a measure of the true effectiveness of the teacher without having to adjust the scores (Miller, 1982). The question then becomes: is the amount of bias in the scores sufficient to warrant adjustment of the scores? This is a complex process (Wigington, Tollefson, & Rodriguez, 1989) which:

Involves much more than determining which instructors earned the higher numerical rating and deciding they are the better teachers. Rather, an interpretation of student ratings needs to reflect an understanding of the variables that interact to produce differences in student ratings of instructors.

(p.343)

3.3 Reliability

Reliability refers to the consistency with which the effectiveness of an instructor is measured. To index reliability, multiple measures of the attribute, teaching effectiveness, are collected on each instructor from multiple independent raters, the students. Certain measures of reliability are used under the theoretical assumption that the construct measured should be stable or consistent under the conditions in which the measure of reliability is taken. However, for student ratings several different measures of reliability may be used to obtain
different results (McKeachie, 1979). The purpose of the evaluation must be taken into consideration before applying these measures, as must factors relating to data collection that may affect the reliability of the scores.

For example, if the same student rates the same instructor at two points in time we hope to obtain a high correlation (.7 or higher) between the two scores. However, we would not want a high correlation between repeated measures of the same instructor if theory suggests that some change is to be expected, particularly if the instructor is expected to improve over time as a result of feedback. If theory or previous experience suggests that different courses have characteristics which might influence the ratings, the correlations between different courses taught by the same instructor may also be low (less than .3).

However, if instructors adjust their teaching to accommodate these course characteristics the correlation between different types of courses may remain high. Factor analysis is often used to confirm the structure of questionnaires where a number of questions are included to measure a single construct (Crocker & Algina, 1986). In this case, correlations between items on the questionnaire that do not measure the same construct should also be low.

There are two types of reliability that are related to this study: internal consistency, which refers to the extent to which items on an instrument are measuring the same construct, and stability, which refers to the consistency of the scores over time. For internal consistency procedures it is assumed that each item on the instrument partially measures the characteristic being assessed and also has a substantial amount of random error. Since random error tends to cancel out when it is averaged across items, the reliability of the mean score will vary with
the number of items and the degree to which the items are cohesive, or measure the same underlying construct.

In cases where scores are obtained by averaging the ratings of individuals, the reliability of the score is affected by the number of items used in the scale (or scales), the internal consistency of the scale, the extent of the agreement between the members of the group, and the number of members sampled per group. Rowen, Raudenbush and Kang (1991) demonstrated that the reliability of group-level scores depends mostly on the number of group members sampled. The number of items on the instrument becomes more relevant if the group sampled is quite small. Results of their analysis showed that reliability tends to increase up to about five items and then plateaus.

Hogan (1973) suggested that scores that were unreliable, such as those obtained from small classes with fewer than ten students, would not be considered acceptable for use in making administrative decisions. For student ratings, Kane, Gillmore & Crooks (1976) found that the negative effect of a small number of student raters on reliability could be offset by increasing the number of items on the survey. For a class size of ten, the generalizability coefficient for the mean of the items was raised from .27 to .58 when the number of items was increased from one to ten. This indicates that in situations where the class size is small, the use of the item mean rather than the global score should provide a more reliable measure.

Crooks and Kane (1981) examined the generalizability of the item mean and the global item, and found that both provided average ratings with high generalizability (coefficients exceeded .8 provided at least 20 students in the section). However they suspected that the mean of the items would pose a validity problem if different items
systematically varied in importance to students in different courses or departments. This would invalidate the reliability index for some classes even if the average of the responses might produce an index with higher generalizability than the global item. Therefore it would appear that an item mean only provides the more reliable measure in cases where the content of the survey is deemed to be valid for the context in which it is administered.

The total number of courses taught by each faculty member may vary considerably. This raises the question of how many courses should be used in the evaluation process to provide a fair estimate of the ability of the instructor. Results from generalizability studies have suggested that a sample of fewer than five courses taught by an individual instructor would be questionable, and that more courses would be required for a dependable score if the instructor taught relatively small sections (Gillmore, Kane and Naccarato, 1979). When more students are surveyed, reliability increases in a systematic way. These researchers suggest that the optimal sample for an instructor would consist of five to ten classes where the average class size is twenty or more students.

The application of internal consistency procedures to student evaluations assumes that each student rater within the same class functions as an individual test item. The reliability of the class-average response is based on all students within the same class (equivalent to the total test score averaged across all test items) and is a function of the relative agreement among different students within the same class (equivalent to the extent to which different test items measure the same underlying characteristics) and the number of students in the class (test items). Differences among students within the same class (all within-class variance) are assumed to be error variance, as are the differences between test items.
If there is systematic variance associated with the ratings of individual students, averaging the responses results in the underestimation of the reliability of a single rater (Marsh & Overall, 1979) and does not allow for any true differences of opinion among students. This also implies that individual student characteristics have no impact on student ratings of teaching effectiveness. Single rater reliabilities are taken from internal consistency measures (agreement among different students in the same class) and stability coefficients (agreement in the ratings of the same students over time). The stability measure will reflect any systematic variance attributable to the individual student responses. If individual students and subgroups with similar characteristics systematically differ, then the source of variation could be identified and this information could be used as a means of matching teaching methods and instructors to a particular student.

Most reliability studies are based on the class rather than the student as a unit of analysis since the difference among teachers is the more important consideration in making employment decisions. If the course that is taught makes a systematic difference in the rating received, reliability estimates will be overestimated for the instructor. Also, if student ratings are more heavily influenced by the type of course taught than by the influence of the instructor, one could argue that poor ratings are caused by unfavorable course assignments rather than by poor teaching performance. This gives rise to studies that examine the stability of student ratings.

Hogan (1973) suggested three questions related to reliability that are relevant to the use of student ratings: 1) How stable are student ratings for the same course taught by the same instructor from semester to semester? 2) How stable are student ratings of an instructor
teaching different courses? 3) How stable are student ratings for the same course taught by different instructors? These questions have important implications for how often a particular instructor teaching a particular course should be evaluated, for course assignments, and for course revision. Short-term stability research includes studies that compare ratings from one term to the next or between mid-term and end-of-term ratings for the same course. Such studies have indicted that students ratings are highly stable (Costin, Greenough & Menges, 1971; Frey, 1978; Kohlan, 1973; Overall; & Marsh, 1979).

Research on ratings indicates that evaluations of a given instructor are reasonably stable across different offerings of the same course, but are much less consistent across different courses or course types taught in the same year. Marsh (1982) compared student ratings of the same course taught by the same instructor on two different occasions (n=34) and found that on average the reliability was high (mean $r = .71$) but lower than the average reliability of the individual courses (mean $r = .93$). Sources of variance included more favorable evaluations being given to courses in which the students expected higher grades, in one course in which the students perceived the course to require the most work, and in one course that was taught after the instructor had taught the course at least once before. Murray (1980) found that reliability coefficients ranged from .62 to .89 (mean $r = .74$) for the same course taught by the same instructor in successive years, compared to reliability coefficients ranging from .33 to .55 (mean $r = .42$) for different courses taught by the same instructor in the same year. Other research showed an average correlation of .71 in the same course taught in different years by the same instructor compared to an average correlation of .52 for different courses taught in the same year by the same instructor (Marsh, 1981). Bauswell,
Schwartz and Purohit (1975) found correlations between the same instructor and the same course somewhat lower (mean $r = .69$), as were the correlations for the same instructor teaching different courses (mean $r = .33$).

Collectively, this research suggests that teaching effectiveness is to some extent context-dependent and that some instructors are more suited to certain types of courses rather than being uniformly effective or ineffective as an instructor for all course types. Hogan (1973) suggested that if ratings for the same course were quite similar on two occasions there may be no need to obtain ratings each time an instructor gives the course unless substantial changes are made in the teaching approach. In addition, if successive ratings are dissimilar, course revisions should not occur on the basis of one set of ratings. It is also important to look at the variety of courses taught when considering instructors for promotion or tenure, because the assumption is often made that it is the instructor who makes the difference in the ratings when the larger influence may be the nature of the course itself.

In a forerunner of the hierarchical model, Kane, Gillmore, and Crooks (1976) suggested an ANOVA design in which students were nested within instructors, crossed with items. The results would yield an instructor effect, an item effect, a student within instructor effect, an instructor-by-item interaction and the residuals. They added a third facet to the design in 1978, and used two designs, instructors nested within courses, and courses nested within instructors. The separate analyses showed considerable variance due to the instructor-by-course interaction, which implied that the performance of an instructor varied substantially from course to course. Because the data collection was limited to two different courses per instructor, this study could not be extended to compare the variability in
performance across occasions for the same course to the variability across different courses. If variability across occasions of the same course is large, the assignment of instructors to a course on the basis of student ratings would not be an effective practice, since a good performance one semester would not necessarily imply a good performance the next semester.

Smith (1979) used a four factor ANOVA design with occasions crossed with instructors, and items and students nested within occasions. He collected data on instructors who taught the same course during three successive semesters. Results showed that the largest variance was with the residual (r = .51) and the instructor-by-student interaction effects (r = .37), with moderate effects sizes for instructor, item, and instructor-by-item interaction effects. Effect sizes for occasions and the instructor-by-occasion interaction were near zero, indicating little or no variation in instructor performance in a specific course from one occasion to the next. They concluded that it would be sufficient to judge a particular instructor on one occasion per course if enough students were included in the sample (n=20, r= .79). For a class size of one, twenty courses would have to be evaluated in order to achieve the same reliability.

The analysis did not include cases where an instructor had no teaching experience or was teaching the course for the first time. Therefore, these results could not be applied to first-time instructors. This suggests that longitudinal research taking the career path of the faculty member into consideration is required in order to effectively and fairly rate all instructors. Previous long-term stability research on student ratings includes studies
concerned with end-of-term ratings and ratings obtained at least one year after graduation, including comparisons of alumni ratings which employ a cross-sectional design.

*Longitudinal Designs*

The stability of teaching behavior can be formulated through two questions: 1) Is the behavior of an individual teacher consistent over time? and 2) Are individual differences among instructors consistent over items (Rogosa, Floden and Willet, 1984). If consistent individual differences in teaching behavior are detected, this creates a link to the detection of individual differences in instructor effectiveness. It is usually assumed that these differences are consistent, and variation is due to measurement error. If differences are found, then it provides information about what kinds of teachers are the most consistent, and in what situations and for what periods of time they are most effective. This logic can be applied to the study of student ratings of instruction. The collection of student ratings of effectiveness can provide a link to the detection of individual differences in teaching behavior. The detection of systematic individual differences in consistency contributes to an understanding of how and why scores fluctuate. Temporal stability (consistency over time) requires that scores remain unchanged over time. However, longitudinal studies could also indicate strong upward or downward trends in the scores and indicate the strength of these trends.

Marsh (1980a) used a within-instructor design based on the assumption that there would be reliable differences in the ratings of the same instructor teaching the same course on two different occasions and found a substantial portion of the variance in the ratings was unique to a particular offering of a given course. He also found that when the same instructor
taught the same course on two different occasions, the ratings tended to be better the second time— instructors apparently get better with practice. However, it was not clear that the practice effect could be generalized to different courses or was specific only to future offerings of the same course. When data collection is limited to two occasions there is no way to determine whether this effect is linear, shows diminishing returns, or perhaps declines over the long term. Two studies of students of evaluations have used longitudinal data.

Marsh and Hocevar (1991) observed that cross-sectional studies provided a weak basis for inferring the future ratings of inexperienced instructors or the prior ratings of more experienced instructors. They used student evaluations from 6024 classes taught by 195 instructors in 31 academic departments collected over a one year period. They found no changes over time on the content-specific factors, the overall course mean, or the global rating and concluded that individual differences in teaching effectiveness were stable. This conclusion supported the use of aggregated ratings across different courses. This study was somewhat limited, in that data were collected in a short period of time and did not take into account variables such as age, rank and experience of the instructor that may covary with time.

Marsh and Bailey (1993) conducted a profile analysis to examine the difference between instructors in their profile of ratings for the different factors, the argument being that the existence of reliable individual differences in profiles of the dimensions would be important, particularly if each instructor had a distinguishable profile from other instructors. To determine if there were individual differences in profiles that generalized across ratings of the same instructor, data were collected over a thirteen-year period. Results showed that
instructors appeared to have distinct profiles of strengths and weaknesses, at least when the scores were aggregated over many students. More variance was attributed to within-instructors, rather than between-instructors, but other course and instructor variables that may have contributed to the variance were not considered in the analysis.

These studies incorporated longitudinal data, but were limited for two reasons. First, the studies incorporated a limited number of possible biases into the study, which is consistent with Marsh’s view that these factors do not contribute to a significant part of the variance in the scores. Second, they fail to include additional variables that may be related to both the instructor and the notion of time, such as age of the instructor, teaching experience at the time the course was taught, and academic rank. Feldman (1983) reviewed related literature and found that these factors showed distinct patterns of relationship to the scores even though they were only weakly related. Evaluations tended to be negatively correlated with age, and to a lesser extent, years of teaching experience. The evaluations were positively correlated with rank. He suggested that the strength of the relationship of ratings with experience and possibly age may be underestimated in studies which explored only a linear relationship, and speculated that a U-shaped relationship occurs peaking somewhere between 3 and 12 years of teaching experience. He also suggested that seniority may covary with other variables that affect ratings. Rank may produce different expectations in students, instructors may change how they teach as they grow older, or different cohorts of students may change their expectations. In studies that used cross-sectional data significant results may have been found because of differences in cohorts of teachers.
Correlational studies provide test-retest information that addresses the question of whether or not differences among instructors are maintained from one occasion to the next, but this information is unsuitable for tracking trends over time. Generalizability coefficients indicate the ability to detect differences among instructors’ average behavior, and give information as to how well the average behavior of an instructor can be located relative to other instructors’ average behavior, but ignore time trends by focusing on the average, construe all variation as error, and assume a steady state for each instructor. The longitudinal models examined in this study will attempt to ameliorate some of these issues in addition to including course and instructor covariates that may affect the scores.

Summary

Three issues relevant to this research have been identified and discussed in this review: 1) the selection of an appropriate outcome measure; 2) the identification and handling of factors that might bias the scores; and 3) the reliability of the scores and long-term stability.

Previous research has indicated that the item mean and the global score may be used interchangeably with modest adjustments, but reliability problems related to sample size affect the choice of score as does the content validity of the items for the raters. Factor scores have little generalizability across different instruments as research into the number and type of factors has been inconclusive and inconsistent. Procedures as to the use of a weighted mean have yet to be explored. However, the selection of the most appropriate score should be related to the contextual factors at the site of the evaluation and the content and construction
of the instrument. Descriptive statistics must also be examined in different evaluation contexts before decisions are made as to the most appropriate outcome measure.

Research into potential biases persists, since the possibility of their influence raises questions about the validity, interpretation and the utility of the scores for administrative decisions. Potential sources of bias include student, course and instructor characteristics. A variety of potential student biases have been identified, with marginal and mixed results. The mixture of study designs which use the ratings at the individual level or aggregate the data at the course level could account for some of the inconsistency, and pooling groups of classes may mask much of the variability which exists from class to class.

Course variables found to have an effect on the scores include class size, course level, course difficulty, required or elective course, grades and the manner of administration of the evaluation. Recent research suggests that the contextual effects of departments and faculties should be taken into consideration when examining the effects of these variables on the scores. Instructor characteristics have not been shown to have consistent relationships with student ratings. There may also be a contextual effect involved with these factors, but it can be argued that many instructor characteristics studied are not really exogenous variables. The instructor variables of interest for score adjustment are those factors which contribute to the variance in the scores but are outside of the control of the instructor.

There has been little research as to the practical outcome of research on biases. Researchers suggest that multivariate analysis techniques are a way to both ameliorate inconsistent findings and examine the cumulative effects of a number of variables that may
influence the scores. It has been suggested that norms or criteria need to be established, but this issue has not been addressed in the literature.

Collectively, the research on reliability suggests that teaching effectiveness is to some extent context-dependent and that some instructors are more suited to certain types of courses rather than being uniformly effective or ineffective as an instructor for all course types. Research indicates that stability estimates are higher when the same instructor teaches the same course rather than different courses, but this research has been limited to courses taught on two occasions and has not, for the most part, taken the opportunity to use the large longitudinal databases that are in existence. This research has also failed to take into account variables that may be related to the scores and may covary with time, such as the age of the instructor, teaching experience at the time the course was taught and academic rank. Time trends are an area of research on student ratings that have yet to be explored.
Chapter Three: Research Methodology

3.1 Subjects and Data Collection

The Faculty of Education at the University of British Columbia offers a variety of programs, including an undergraduate degree in Human Kinetics, post baccalaureate teacher education and diploma programs, and graduate degrees at both the masters and doctoral level. From 1982 to 1993 student evaluations of teaching were administered in classes throughout the faculty. The Standing Committee on the Evaluation of Teaching (SCET) was responsible for the design, administration and analysis of these evaluations, and the maintenance of an archival database from which the sample for this study was drawn.

The questionnaire used was a computer scanable form on which students were asked to provide demographic information, fill in responses to 30 items related to effective teaching and respond to a global question pertaining to the overall quality of the teaching. Responses to the 30 items were obtained using a 7-point rating scale ranging from (1) disagree very strongly to (7) agree very strongly. Responses to the global question were obtained using a 7-point rating scale ranging from (1) very poor to (7) excellent. The same questionnaire was used consistently for the time period under study. A copy of this questionnaire is provided in Appendix 1.

Out of 6,266 courses evaluated, only those taught by tenure-track or tenured assistant, associate and full professors were selected. Classes taught by sessional instructors, teaching assistants and regular instructors were excluded since the policy issues related to the use of ratings are somewhat different for these groups than for the tenure-track and tenured
Because a relatively small number of the classes selected had an enrollment of greater than 100 students (n = 15), these classes were treated as outliers and excluded from the analysis. The final sample consisted of 3,689 courses taught by 260 instructors.

The instructor sample consisted of 67.3% (n = 175) males and 32.7% (n = 85) females. In 1993, 29.2% (n = 76) of the instructors were full professors, 36.5% (n = 95) were associate professors, and 32.2% (n = 171) were assistant professors.

3.2 Measures

Dependent Variables

After the completion of the course and the submission of final grades individual instructors and administrators were provided with a summary of the results of the student evaluations. These summaries included the class means for each item, the mean of all 30 items, and the class mean for the global item. Two outcome measures, the mean of all 30 items on the evaluation (Composite Score) and the response to the Global Item, have been discussed in the previous chapter as possible scores for use in administrative decisions. The sample mean for the Global Item was 5.82 (sd = 0.81), and the sample mean of the Composite Score was 5.75 (sd = 0.74). Although a significant correlation of .94 between the two means indicated that the two scores represented the same construct, results of a paired t-test showed that the difference between the two means was significant (t = -14.67, p < .000). To examine this difference both measures were included in the initial part of the analysis.
Independent Variables

The design and purpose of the research necessitated the inclusion of two sets of independent variables: 1) characteristics of the individual courses; and 2) characteristics of the instructors who taught those courses. The first set included class size, percentage of females in the course, percentage of students taking the course as an elective, level of the course, the date of the evaluation and the years of experience of the instructor who taught the course. These variables were used in the first level of the model in which within-instructor variation was modeled. The second set consisted of sex and rank of the instructor and the total number of courses each instructor had taught. These variables were used in the second level of the model in which between-instructor variation was examined.

Class size was entered into the database as the number of students who completed the questionnaire. These values were centered on the mean and divided by 100 for the analysis. The resulting values were then squared and cubed in order to fit and test a polynomial relationship between the outcome score and class size. These variables were labeled SIZE, SIZESQD and SIZECBD.

The student responses to sex were coded (males = 0, females = 1), and the percentage of females in each class was calculated. These values were then centered around the mean. Similarly, students were asked to indicate whether they were required to take the course or were taking it as an elective. These responses were coded (required = 0, elective = 1), and the percentage of those taking each course as an elective was calculated and centered around the mean. These variables were labeled PCTFEM and PCTELEC.
Courses taught in the faculty are numbered from one hundred to six hundred. Courses at the 500 and 600 level are graduate courses. Courses at the 400 level are postbaccalaureate education courses and advanced diploma courses. Courses at the 300 level are a mixture of undergraduate, junior diploma, and entry-level teacher education courses, while 100 and 200 level courses are primarily undergraduate courses taught in the school of Human Kinetics. The two levels of graduate courses were combined into one, and the 400 level courses were kept as a level by themselves. Means for the 100, 200 and 300 level courses were similar, therefore these three levels were combined. In addition, the relatively small number of first and second year courses (n = 158) in relation to the sample justified their inclusion in this category. The three variables created, LEV123, LEV4 and LEV56, were dummy-coded and centered on their respective means for the analysis.

The variable TIME represents the date of the course evaluation. If a course was evaluated in January of 1982 it was given a value of 1982. If the course was evaluated in January of 1983 it was given a value of 1983. The interval between (one year) was divided by 11 to provide a value for each subsequent month. So a class that was evaluated in February of 1982 would have a value of 1982.091, March of 1982, a value of 1982.183, and so on for each subsequent month and for every year in the database. The first year in the database, 1982, was subtracted from each point in time to create values ranging from .498 to 11.664. These values for TIME were then squared and cubed (TIMESQD, TIMECBD) in order to fit and test the possibility of a polynomial relationship for one of the longitudinal models.

For the variable *years of experience of the instructor who taught the course*, five variables were created from the data. Initially each course was coded as to the years of
experience the instructor had at this university at the time of the course. Then each course was coded into the appropriate period in the career path of the instructor. Conceptually, the career path of an instructor was divided into five segments. Period one (PER1) included the first five years of the instructors' career, normally the time during which they would be working toward achieving tenure. Period 2 (PER2) included the next five years of the career path at the end of which there would be a review, and the possibility of an increase in rank. Period three (PER3) represented the next five years, normally ending with a review and possible promotion. Period four (PER4) included the next ten years of experience, and period 5 (PER5) the final ten years in the careers of faculty included in this database.

At the instructor level, the number of courses taught by each instructor were totaled and used as a proxy for experience and workload (NUMCRSE). The sex of each instructor was coded (males = 0, females = 1) and entered into the database under the name SEX. The highest rank obtained by each instructor within the time-frame of the database was determined and entered into the database. Three dummy-coded variables were used: RANK0, for assistant professors; RANK1, for associate professors; and RANK2, for full professors.

### 3.3 Data Analysis

*Hierarchical Linear Models*

This analysis used a statistical technique known as hierarchical linear modeling or HLM (see Bryk & Raudenbush, 1992; Bryk, Raudenbush, Seltzer, & Congdon, 1989). This technique was developed to enable researchers to use a multilevel regression analysis on data that fits an hierarchical model. The strength of an hierarchical analysis is that it estimates
statistics for each unit of an hierarchical structure using data for that unit, while borrowing strength from the information available on all units. In this study there were two units of analysis: the course and the instructor. The data fit an hierarchical model with the courses nested within the instructors who taught them. A statistical explanation of the technique is not provided here but can be reviewed in detail in other sources (e.g. Bryk & Raudenbush, 1992; Lee & Bryk, 1989; Rumberger & Willms, 1992).

This type of analysis allows the variation in a variable (the score on the student evaluation) to be partitioned into within-instructor and between-instructor components. It also allows the relationship of the score and variables from both the within-instructor and between-instructor levels of the analysis to be examined. An important element of a multivariate analytic model is that characteristics that might confound the relationships of interest are taken into account. Since the objective of this analysis is to separate the effects of exogenous variables from the “true” effectiveness of the teacher, it is important to evaluate the effects of instructors on the responses of students after controlling for course characteristics.

The hierarchical linear model used in this study has two levels: courses and instructors. The first level examines the relationship between the course evaluatee’s scores (Composite Score or Global Item) and course characteristics (i.e. class size, percentage of females in the course, percentage of students taking the course as an elective, level of the course, the date of the evaluation and the years of experience of the instructor who taught the course). The analysis essentially estimates this regression model separately for each of the 260 instructors in the sample. Thus there are 260 separate estimates of the intercept, which
indicate an ‘instructors’ average evaluation score after adjusting for the course-level variables in the model. At the second level these adjusted means (i.e., the intercepts from the level-1 regressions) become the dependent variable for a between-instructor regression. The analysis then examines whether the adjusted means are related to instructor characteristics such as sex, rank, and total number of courses each instructor had taught.

The first level regression also produces 260 estimates of each of the other regression parameters (independent variables). HLM provides a test of whether the variation in the parameter estimates is larger than what would be expected if it were simply random variation. For example, the level-1 analysis estimates within-instructor parameters of interest such as class size and course level which may be set to be fixed or to vary randomly across instructors. With the exception of the variables date of the evaluation and years of teaching experience, all course-level variables were centered on their means, a process known as “centering”. This computation does not affect the estimated values for the regression slopes, but does change the value of the intercept. The resulting intercept is an estimate of what the mean achievement for each instructor would be if the characteristics of the courses taught by that instructor were the same as the average characteristics for the sample. As a result of centering, the instructors can be compared in terms of how effective they would be if they had taught courses with similar characteristics.

The between-instructor variation in the random parameters is investigated in the second stage of HLM. In this step, instructor characteristics are identified that may show systematic relationships with the within-instructor variables. At this level of the analysis, the
instructor-level model attempts to explain variation in the estimated coefficients for each instructor from the course-level model with instructor-level explanatory variables.

Considerable research has employed this type of model in the examination of school effects (see Lee & Bryk, 1989; Raudenbush & Willms, 1995), but this technique has not been employed in the study of student evaluations of teaching. HLM addresses both the hierarchical nature of these data and the research questions for this study, and also takes into account the variability in the number of courses taught by each instructor. When the parameters are estimated, HLM weights these estimates by the precision with which they were estimated (Bryk & Raudenbush, 1992, p.34ff.). When the reliability of a statistical parameter is influenced by within-instructor sample sizes, HLM takes this variable reliability into account so that particularly large sample sizes do not overwhelm the analysis. Thus it allows all of the data to be used with confidence since the variation in within-instructor sample size is accounted for in an unbiased fashion.

Models in the Analysis

The first model of the analysis is called a null model because it does not include the independent variables used for course and instructor characteristics. This model is used to partition the variance in the outcome measure into within-instructor and between-instructor components. If all of the variance were attributable to only within-instructor variation, the examination of a two-level model would be unnecessary. In this case, multiple regression could be used to examine the effects of the independent variables on this variation.
However, if there is variation to be explained between instructors, and further analyses are performed, the null model provides a baseline from which to calculate the amount of variance that is explained by subsequent models. The null model was run for both the Composite Score and the Global Item. In addition, a succession of null models based on the Composite Score as the outcome measure and the number of courses taught per instructor were run to determine the number of classes necessary to obtain a reliable score to be used for personnel decisions.

The null model was then extended to include course characteristics which might influence the scores. This within-instructor model was estimated using the variables *class size, percentage of females in the course, percentage of students taking the course as an elective* and *level of the course*. The estimated coefficients provide a measure of the difference in scores that are based on these variables. This allows the researcher to find out if there are statistically significant differences in the estimated coefficients among instructors. It also provides an estimate of the proportion of variance that is explained by the model. This model was then extended to two levels to include the instructor characteristics *sex, rank, and total number of courses each instructor had taught*. This model attempted to explain the variance in any of the random level-1 course characteristics with these variables.

The fourth and fifth models were designed as longitudinal models, each utilizing a different conceptualization of time. In the fourth model TIME, TIMESQD and TIMECBD were used as course covariates in order to investigate a growth model. This model would result in an average picture of how the scores changed or did not change over the selected chronological period of time. In studies of student ratings the data have typically been
collected at only two time points. Such designs are often inadequate for studies of individual growth (Bryk & Raudenbush, 1987; Rogosa, Brandt & Zimowski, 1982). In this study, multiple observations were collected on each instructor and these observations were seen as nested within the instructor. The nesting allows the researcher to carry out this type of analysis even when the number and spacing of time points vary across instructors (Bryk & Raudenbush, 1992). In the first level of the model, each instructors’ development is represented by an individual growth trajectory that depends on their unique set of parameters. These individual growth parameters become the outcome variables in the second level of the model, where they may depend on some instructor-level characteristics.

In the fifth model time was conceptualized as following the career path of the instructor. The years of experience variables which were created and labeled PER1 through PER5 were used as covariates at the course-level. Each course was coded as to the period of experience in the career path the instructor was in at the time at which the course evaluation occurred. The model was designed as a piecewise linear growth model (Bryk & Raudenbush, 1992). This type of longitudinal model is used in cases when the data suggest nonlinearity and when the researcher wishes to compare growth rates at different time periods. In this case, the model was designed to explore the differences in scores at different time periods in the career paths of the faculty. It also provides a normative model of the average scores over the career path of an instructor which can then be compared with the actual longitudinal profiles of scores for individual instructors.
Chapter Four: Results

4.1 Descriptive Statistics

Table 1 shows the means and standard deviations for the Composite Score and the Global Item. The Global Item was consistently greater than the Composite Score across all course levels, ranks, and the sex of the instructor. The difference between the two sets of scores ranged from .04 to .08, which represents less than 10% of a standard deviation for each variable. The standard deviation was larger for the Global Item in all cases, which indicates that these scores were less homogeneous.

<table>
<thead>
<tr>
<th></th>
<th>Composite Score</th>
<th>Global Item</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean</td>
</tr>
<tr>
<td>Average Score</td>
<td>3689</td>
<td>5.74</td>
</tr>
<tr>
<td>Course Level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEV123</td>
<td>1490</td>
<td>5.63</td>
</tr>
<tr>
<td>LEV4</td>
<td>994</td>
<td>5.67</td>
</tr>
<tr>
<td>LEV56</td>
<td>1206</td>
<td>5.94</td>
</tr>
<tr>
<td>Rank of Instructor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assistant professor</td>
<td>1171</td>
<td>5.65</td>
</tr>
<tr>
<td>Associate professor</td>
<td>1525</td>
<td>5.74</td>
</tr>
<tr>
<td>Full professor</td>
<td>1006</td>
<td>5.84</td>
</tr>
<tr>
<td>Sex of Instructor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>2553</td>
<td>5.73</td>
</tr>
<tr>
<td>Female</td>
<td>1149</td>
<td>5.79</td>
</tr>
</tbody>
</table>
On average female instructors were rated marginally higher than male instructors, and their scores were slightly more variable for both the Composite Score and the Global Item. For both scores, ratings increased as the level of the course increased and as the rank of the instructor increased. The difference between the average scores of lower level courses and graduate level courses represents about 42% and 33% of a standard deviation for the Composite Score and the Global Item respectively. The difference between the scores for an average assistant professor and full professor represents about 25% of a standard deviation for the Composite Score and 21% of a standard deviation for the Global Item.

The variation of both scores decreased as the level of the course increased and also decreased as the rank of the instructor increased. This indicates that scores for higher level courses and full professors were more homogeneous. However, the difficulty in comparing standard deviations from these results is that they include both sampling error and measurement error. In the hierarchical regression models which follow, it is possible to partition the variance and provide an estimate of the “true” variance between instructors for the ratings.

Means and standard deviations for the course and instructor variables (before centering) are shown in Table 2. Class size ranged from 1 to 98 students. At the ends of the distribution, 2.6% of the courses (n = 96) had an enrollment of less than 5 students, while 2% of the courses (n = 78) had an enrollment of 45 to 98 students. The years of experience of the instructors ranged from less than one year to 35 years of experience. The number of courses in the sample taught by an individual instructor ranged from 1 to 70. Many of the faculty had
instructed 5 or fewer courses (21.2%); fewer faculty had instructed more than 29 courses (10%).

Table 2
Descriptive Statistics for Course and Instructor Variables

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class size</td>
<td>17.39</td>
<td>11.34</td>
</tr>
<tr>
<td>Percentage taking course as an elective</td>
<td>34.19</td>
<td>35.32</td>
</tr>
<tr>
<td>Percentage of females in the class</td>
<td>67.35</td>
<td>24.09</td>
</tr>
<tr>
<td>Time</td>
<td>87.89</td>
<td>3.32</td>
</tr>
<tr>
<td>Years of experience at the time course taught</td>
<td>14.42</td>
<td>7.78</td>
</tr>
<tr>
<td>Number of courses taught</td>
<td>14.24</td>
<td>10.28</td>
</tr>
</tbody>
</table>

4.2 Null Models

Table 3 shows an analysis of the variation in scores among instructors for both the Composite Score and the Global Item. These are the results of a null model, which is identical to a random effects analysis of variance which separates the variance into within- and between-instructor components. In this model no course- or instructor-level factors are used to explain variation in instructors' ratings. The model provides estimates of the mean score for each instructor, and an estimate of the precision-weighted mean of the instructor means.

The analysis also estimates the standard deviations of the course means between- and within-instructors for each outcome measure and tests their statistical significance. If the standard deviation between instructors is not statistically significant, then the observed variation between instructors in their ratings is probably due to measurement and sampling
error. If that were the case, then further analyses to determine the sources of these differences would be unnecessary.

In these models, the estimated standard deviations (.467 and .525) were both statistically significant. Usually 95% of the estimated instructor means would lie within plus or minus two standard deviations of the sample mean. For the Composite Score, the instructor means ranged from approximately 4.79 to 6.65 (5.72 ± 2(.467). This is a range of 1.87 SDs, which is equivalent to about 1.38 points on the 7-point rating scale (i.e., 1.868 times .74, the SD of the Composite Score reported in Table 1). Similarly the range for the Global Item is 2.10 SD’s, which is equivalent to about 1.70 points on the 7-point rating scale. These results show that there was a substantive amount of variation among instructors on the ratings.

Table 3
HLM Results for the Null Models of the Composite Score and the Global Item

<table>
<thead>
<tr>
<th></th>
<th>Composite Score</th>
<th>Global Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of Average Ratings</td>
<td>5.72</td>
<td>5.79</td>
</tr>
<tr>
<td>Standard Deviation (Between)</td>
<td>.467</td>
<td>.525</td>
</tr>
<tr>
<td>Standard Deviation (Within)</td>
<td>.574</td>
<td>.622</td>
</tr>
</tbody>
</table>

Model Statistics

<table>
<thead>
<tr>
<th>Components of Variation as Percentages</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Within Instructors</td>
<td>60.2</td>
<td>58.3</td>
</tr>
<tr>
<td>Between Instructors</td>
<td>39.8</td>
<td>41.7</td>
</tr>
<tr>
<td>Reliability of the Estimates</td>
<td>0.84</td>
<td>0.85</td>
</tr>
</tbody>
</table>
For the Composite Score, the estimated variance of the courses taught by an instructor was .330 and the variance between instructors was .218. When these values are expressed as percentages, 60.2% of the variance in the Composite Scores is between courses within instructors, and 39.8% of the variation is between instructors. For the Global Item, 58.3% of the variance in scores is between courses within instructors and 41.7% of the variation is between instructors.

These findings show that there is considerable variation to be explained both within and between instructors, although instructors varied more on the courses they taught than they did when compared with other instructors. Use of these scores without taking the within-instructor variance into account implies that differences in scores between instructors are due to the quality of instruction alone, and ignores differences in course and instructor characteristics that might influence the scores and account for some of the variance. For normative and comparative purposes, examining causes of variation which are not a function of the quality of instruction becomes important, particularly if variation among instructors can be explained beyond measurement and sampling error by factors external to instruction.

The final line in Table 2 reports the overall reliability with which the average scores of the instructors can be estimated. The reliability of the estimate depends on the accuracy with which the indicators are estimated, and the magnitude of the ‘true’ variance of the scores. Both scores provided good reliability (.84 and .85), indicating that the sample means were quite reliable as indicators of the true instructor means.

The conclusion drawn from this analysis is that there was substantial variation in the scores within- and between-instructors, even after taking into account measurement and
sampling error. In reports which provide only the scores, the implication is that differences in the scores are due to differences in instruction alone, ignoring within-instructor variation and the effects of course and instructor characteristics. The results of the null model provide the foundation for examining reasons for the variation within- and between-instructors. The results of these analyses are presented in section 4.3.

The results of the null model for both outcome measures were quite similar and the differences were of little practical significance. Since the Global Item was more variable it may have been a better candidate for further analyses, but preliminary results of subsequent models showed results similar to that of models using the Composite Score. Because the sample included courses with fewer than ten student raters, the Composite Score was used to offset any negative effects on reliability due to small sample sizes. Therefore the remainder of the analyses in the study used the Composite Score as the outcome measure.

The second part of the null model analysis involved successive runs of the model with the number of courses taught as the independent variable and the Composite Score as the outcome measure. In the first run, the first course in the sample taught by each instructor was included in the analysis. In the second run, the first two courses taught by each instructor were included, and instructors with only one course were excluded from the analysis. In the third run, the first three courses taught by each instructor were included, and instructors with fewer than three courses were excluded. This procedure was followed until each of the remaining instructors had taught a total of 26 courses. The reliability coefficients obtained from each of the models are plotted in Figure 1. These results showed that fewer courses were required for a reliable rating than previous research had indicated.
Many authors agree that a standard reliability level of .70 is reasonable for individual student evaluations (Feldt & Brennan, 1989). This standard may not be sufficiently stringent when important decisions are to be made based on the results of these evaluations, since "high reliability is required when the decision is important, final, irreversible, unconfirmable, concerns individuals and has lasting consequences" (Gronlund & Linn, 1990, p. 101). Results of these analyses showed that four classes were required in order to achieve a reliability of .73. Six courses were required to achieve a reliability of .81, which was approximately the level of reliability for the entire sample (.84). It was not until 16 classes had been evaluated that a reliability of .90 was achieved, and a further 10 courses were required before the reliability leveled off at about .96. This has important implications for the use of SET ratings for tenure and promotion decisions.

Fig. 1 Reliability of SET Scores
In this sample, many instructors in the earlier stages of their careers had not taught many classes. Some faculty members had administrative duties or research activities which resulted in lighter than average teaching loads. In this sample 12.7% of the faculty had not taught the four classes required to achieve a minimal standard of reliability. One fifth of the sample (21.2%) had taught fewer than six courses, and would therefore fall short of the number required for a standard of .80. However, since 63% of the faculty had taught fewer than sixteen courses, a standard for reliability set at .90 would be too high for most members of the faculty. These results raise the question of how much weight can be given to these evaluations if the faculty member does not have a substantial teaching load. It also brings into question the use of the ratings in comparative situations. Can the ratings of an instructor who has taught many courses legitimately be compared to the ratings of an instructor who has taught only three or four courses?

This information about the reliability of the scores can be used when scores for an individual are compared to a predetermined criterion score. In this sample the mean was 5.74 and the SD was .74. We might decide that an average of an individual instructors’ scores should lie within at least one standard deviation of the mean. Therefore, the lower limit of acceptable scores would be 5.00. Suppose an instructor had taught four classes, and the average score for these classes was 4.60. The information from the null model for four classes can be used to calculate the standard error of measurement, and a confidence level for this instructor.

Using the formula $SE = SD\sqrt{1-r}$, we obtain a standard error of .24 (.46$\sqrt{1-.73}$). This indicates the amount of error that must be considered in interpreting the scores for an
instructor who has taught only four classes. It provides the limits within which we can reasonably expect to find the instructors’ “true” rating. To set the confidence intervals, 68% of the obtained scores would fall within plus or minus 1 standard error (.24) of the true score, 95% would fall within plus or minus 2 standard errors (.48), and 99.7% would fall within plus or minus 3 standard errors (.72). These limits can be applied to the instructors’ obtained score of 4.60 to obtain the ranges within which we could be reasonably sure to find the instructors’ true score (4.36-4.84; 4.12-5.08; 3.88-5.32). We can be quite confident that the true score is somewhere between 4.12 and 5.08 because 95% of the observed scores fall within 2 standard errors of the true score. Thus since the upper limit contains the criterion score of 5.00 we can be reasonably certain that the scores for the instructor fall within an acceptable range. However, the use of 1 standard error of measurement is more conservative, and is more common in interpreting individual test scores (Gronlund & Linn, 1990). In the case described above, the average score for the instructor would not be contained within a range that also included the criterion if the range for 1 standard error of measurement were used.

As the number of courses increases, the standard error of measurement decreases. When the standard error of measurement is small, the confidence band narrows indicating higher reliability, and greater confidence that the obtained score is near the true score. For these data, the standard error of measurement for an instructor with 6 courses was .21 (.48\sqrt{1-.81}), which makes it more difficult for an instructor with a low score to fall within a range that contains the acceptable criterion score. The smaller standard error results in a smaller range of acceptable scores, but there will be greater confidence that the obtained
score is a dependable measure of the instruction. When scores are interpreted as bands of scores (confidence bands) less emphasis is placed on the interpretation of minor differences in the scores.

4.3 Course and Instructor Models

Level-1 Model

The next analysis attempts to explain variation in the Composite Score with the course characteristics class size (SIZE, SIZESQD, SIZECBD), percentage of females in the course (PCTFEM), percentage of students taking the course as an elective (PCTELEC) and level of the course (LEV123, LEV4). Table 4 shows the results of this analysis. The hierarchical analysis estimates a separate regression model for each instructor with regression estimates which show the "average within-instructor equation" across 260 instructors. The intercept for each within-instructor equation is the expected average student rating for a course that has average characteristics. In this case the intercept represents the average score for an instructor teaching a graduate course in which 34.19% of the students were taking the course as an elective, 67.35% of the students were females and the course had an enrollment of 17.39 students.

The percentage of students taking the course as an elective and the percentage of females taking the course both had a significant effect. The reported coefficient for PCTELEC can be interpreted to mean that for every 1% increase in the number of students taking the course as an elective the predicted score would increase by .003. Thus, for a course in which all students were enrolled as an elective, the expected score would be 5.86 (5.66 +
Table 4
HLM Analysis of the Effects of Course-Level Variables
on Course Means Within Instructors

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Composite Score</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>SE</td>
<td></td>
</tr>
<tr>
<td>Average Within Instructor Equation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>5.658**</td>
<td>0.032</td>
<td></td>
</tr>
<tr>
<td>PCTELEC</td>
<td>0.003**</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>PCTFEM</td>
<td>0.001*</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>LEV 123</td>
<td>-0.159**</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>LEV4</td>
<td>-0.127**</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td>SIZE</td>
<td>-1.500**</td>
<td>0.149</td>
<td></td>
</tr>
<tr>
<td>SIZESQD</td>
<td>6.522**</td>
<td>1.012</td>
<td></td>
</tr>
<tr>
<td>SIZECBD</td>
<td>-7.196**</td>
<td>1.394</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effects</th>
<th>SD</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.427</td>
<td>0.182**</td>
</tr>
<tr>
<td>PCTELEC</td>
<td>0.004</td>
<td>0.000**</td>
</tr>
<tr>
<td>LEV 123</td>
<td>0.436</td>
<td>0.190**</td>
</tr>
<tr>
<td>LEV4</td>
<td>0.406</td>
<td>0.163**</td>
</tr>
<tr>
<td>level-1</td>
<td>0.514</td>
<td>0.264</td>
</tr>
</tbody>
</table>

Model Statistics

<table>
<thead>
<tr>
<th>Reliability of Estimates</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.789</td>
<td></td>
</tr>
<tr>
<td>PCTELEC</td>
<td>0.337</td>
<td></td>
</tr>
<tr>
<td>LEV123</td>
<td>0.452</td>
<td></td>
</tr>
<tr>
<td>LEV4</td>
<td>0.433</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percent Variance Explained</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Within Instructor</td>
<td>19.8</td>
<td></td>
</tr>
<tr>
<td>Between Instructor</td>
<td>16.6</td>
<td></td>
</tr>
</tbody>
</table>

* p < .05
** p < .01
.003(100 - 34.19) compared with 5.56 (5.66 + .003(0 - 34.19)) for a course in which all students were enrolled as a requirement. The difference between the two extremes (.3) represents about 70% of a standard deviation. This is a large effect that is both significant and substantive.

For every 1% increase in the number of females in the class the scores would increase by .001. Therefore in a course with an all female enrollment the expected score would be 5.69 (5.66 + .001(100 - 67.35)) compared with an expected score of 5.59 (5.66 + .001(0 - 67.35)) for a course in which no females were enrolled. This difference (.1) represents about 23% of a standard deviation. This is a moderately large effect size; however, very few classes in the faculty had an enrollment that was less than 50% female, so the extremes employed provide an effect size that is larger than the general case.

The results were consistent with previous research as to the effect of course levels on the ratings. The score for the intercept (5.66) represents the average score for a graduate course. If an instructor taught a level 100, 200 or 300 course (LEV123), their expected score (5.50) would be -0.159 lower on the 7-point scale than the expected score for a graduate course. If an instructor taught a level 400 course (LEV4) their expected score (5.53) would be -0.127 lower on the 7-point scale than the expected score for a graduate course. These scores differ from the score for a graduate course by about 37% and 28% of a standard deviation respectively. However, the absolute difference between the expected scores for the two levels of undergraduate courses indicates that these two levels could be combined in future research on this faculty.
The results show that the effect of class size is significant even after controlling for the effects of course level, percentage of females and percentage of students taking the course as an elective. In a manner similar to the use of polynomials in simple multiple regression (see Pedhauzer, 1982), the three variables for class size are interpreted as a unit. Taken together, the coefficients indicate that there is a definite quadratic polynomial trend in these data. Figure 2 demonstrates the nature of this curve.

There is a downward trend in the estimated scores as the enrollment increases from 0 to 20 students. For enrollments ranging from 20 to 70 students the scores remain relatively constant. There is a final downward trend for larger courses with 70 to 100 students. These trends become important when viewed in the context of the distribution of enrollments for courses in this faculty. In the sample, 69.3% of the courses taught had an enrollment of 20 or
fewer students, 29.9% had an enrollment of 20 to 70 students and the remaining 0.8% had an enrollment of over 70 students. Recalling that the average class size for the sample was 17.39 students, the question then becomes: which ratings should benefit from adjustments based on class size?

In the lower range of courses, a class size of 5 would have a predicted score of 6.49
\[5.66 + [(17.39-5/100)(-1.5) + (17.39-5/100)^2(6.5) + (17.39-5/100)^3(-7.2)]\]. A class size of 20 would have a predicted score of 5.63, a score very close to the estimate for the average class size in the sample (5.66). The expected score for a class of 20 students and a class of 70 students would be approximately the same. The difference in the expected scores for classes of 5 and 20 students is quite large, (0.86) representing 2.01 standard deviations. Because this is such a large difference, it would seem that some adjustment of the scores is required.

Typically scores for class sizes of less than 5 are not reported, so the difficulty is in deciding whether to adjust scores for larger classes up to the level of the estimated score for a class size of 5 or to adjust smaller courses down to the level of the estimated score for a class size of 20. Another solution would be to adjust the scores for courses that have an enrollment lower than the sample mean down, and adjust the scores for courses that have an enrollment larger than the sample mean up to adjust for the effects of class size.

Contrary to previous research, the trend for larger courses was significant, even though the largest class size in the sample was less than 100. On average, scores for larger classes were lower. These results were inconsistent with previous research which indicated that the relation ship between the scores and class size produced a U-shaped distribution, with small and large classes receiving higher scores. The predicted score for a class of 98
would be 4.80, 3.96 standard deviations lower than the predicted score for a class of 5 and 2.00 standard deviations lower than the average predicted score for the sample. These results indicate that there should be some adjustment in the ratings for classes with very large enrollments. However, the number of classes with an enrollment larger than 70 represents a small portion of the sample. Therefore these coefficients should be used for score adjustment with caution. There may be some other variable such as type of course or level of course that is a more appropriate predictor of this downward trend in scores for large classes than the actual class size.

The second part of Table 4 shows the estimates of the random effects and the significance of these effects across instructors. The variance of the intercepts was statistically significant (p < .01) which means that there were significant differences between instructors on their scores even after controlling for the course level variables in the model. The standard deviation of the adjusted mean is only marginally smaller than that of the null model (unadjusted mean) which indicates that adjustment for course variables has not substantively reduced the variation in scores between instructors. Because the variation in the adjusted score is similar to that of the null model, the reliability of the adjusted measure is reasonably close to that of the unadjusted model. Therefore we are just as able to distinguish between particular instructors in their background-adjusted outcome scores as we were with their unadjusted scores.

The results showed that instructors also varied significantly for the percentage of students taking the course as an elective and on undergraduate courses at both levels. The significant variances indicate that there is a significant effect on effectiveness for some
instructors if the course is an elective, and that some instructors are more effective for different course levels. The reliabilities of these estimates are modest (.34 to .45) indicating a fairly good ability to distinguish between instructors according to the strength of their association on these variables.

The lower portion of the table shows the percentage of within- and between-instructor variation explained by the course characteristics in the model. The partitioning of variance in Table 1 indicated that 60.2% of the variation in student ratings was within instructors and 39.8% of the variation was between instructors. The course characteristic variables in Table 4 accounted for 19.8% of the variation within-instructors and 16.6% percent of the variation between-instructors. The course characteristics explain a substantial proportion of both the within- and between-instructor variation, 18.53% in total. This leaves 81.47% of the total variation left to be explained (48.28% within-instructors and 33.19% between-instructors). Further explanation of the variance is attempted in the next section, which introduces instructor characteristics into the model.

*Level-2 Model*

The unexplained variation between instructors is of interest to administrators who wish to use the scores to compare instructors. Therefore the model was extended to determine if this variation could be partially explained by the instructor factors, *number of courses, sex, rank*. The number of courses was not significant and was excluded from the analysis. This lack of significance showed that on average, instructors who taught more frequently did not differ significantly on their scores from instructors who taught fewer courses. Two models
Table 5
HLM Analysis of the Effects of Instructor-Level Variables on Course-Level Means Within Instructors

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effects</td>
<td>Coefficient</td>
<td>SE</td>
</tr>
<tr>
<td>Average Within Instructor Equation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male Full Professor</td>
<td>5.651**</td>
<td>0.055</td>
</tr>
<tr>
<td>Female</td>
<td>0.120</td>
<td>0.064</td>
</tr>
<tr>
<td>Assistant Professor</td>
<td>-0.043</td>
<td>0.074</td>
</tr>
<tr>
<td>Associate Professor</td>
<td>-0.046</td>
<td>0.070</td>
</tr>
<tr>
<td>PCTELEC</td>
<td>0.003**</td>
<td>0.001</td>
</tr>
<tr>
<td>PCTFEM</td>
<td>0.001*</td>
<td>0.001</td>
</tr>
<tr>
<td>LEV 123</td>
<td>-0.157**</td>
<td>0.049</td>
</tr>
<tr>
<td>LEV4</td>
<td>-0.126**</td>
<td>0.047</td>
</tr>
<tr>
<td>SIZE</td>
<td>-1.153**</td>
<td>0.149</td>
</tr>
<tr>
<td>SIZESQD</td>
<td>6.505**</td>
<td>1.014</td>
</tr>
<tr>
<td>SIZECBD</td>
<td>-7.167**</td>
<td>1.310</td>
</tr>
<tr>
<td>Random Effects</td>
<td>SD</td>
<td>Variance</td>
</tr>
<tr>
<td>Male Instructor</td>
<td>0.426</td>
<td>0.182**</td>
</tr>
<tr>
<td>PCTELEC</td>
<td>0.004</td>
<td>0.000**</td>
</tr>
<tr>
<td>LEV 123</td>
<td>0.441</td>
<td>0.194**</td>
</tr>
<tr>
<td>LEV4</td>
<td>0.406</td>
<td>0.165**</td>
</tr>
<tr>
<td>level-1</td>
<td>0.514</td>
<td>0.264**</td>
</tr>
<tr>
<td>Model Statistics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reliability of Estimates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male Instructor</td>
<td>0.791</td>
<td></td>
</tr>
<tr>
<td>PCTELEC</td>
<td>0.337</td>
<td></td>
</tr>
<tr>
<td>LEV123</td>
<td>0.456</td>
<td></td>
</tr>
<tr>
<td>LEV4</td>
<td>0.433</td>
<td></td>
</tr>
<tr>
<td>Percent Variance Explained</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Instructor</td>
<td>19.9</td>
<td></td>
</tr>
<tr>
<td>Between Instructor</td>
<td>16.8</td>
<td></td>
</tr>
</tbody>
</table>

** p < .01  
* p < .05
are presented which investigate the effects of sex and rank.

In a 2-level hierarchical model, the random effects from level-1 are regressed on the level-2 characteristics (in this case rank and sex). The choice of which random effects to model is in part driven by theory, and is determined by the reliability with which the random effects were estimated. Modeling of the intercept is of primary importance. Table 5 shows the results of an analysis with the intercepts from level-1 regressed on rank and sex (Model 1). Because both rank and sex were dummy coded, the intercept is interpreted as a male instructor at the rank of full professor who would have an expected score of 5.65. When both rank and sex were modeled on this intercept the results were not significant at the .05 level, but the results were of substantive interest. The coefficients showed that the estimated score for a female full professor would be higher than the estimated score for a male full professor (5.65 + 0.12 = 5.77), as would be the estimated scores for female assistant and female associate professors (5.73, 5.72). The estimated scores for male assistant and male associate professors would be slightly lower than the estimated score for a male full professor (5.61, 5.60).

Although these results are not statistically significant or conclusive they do indicate that rank and sex of the instructor do have a subtle effect on the variation in ratings between instructors for this particular faculty. However, inclusion of these variables in the model does not result in an increase in the percent variance explained. The purpose of their inclusion in the model is to determine if there is significant and substantive bias in the scores that can be attributed to these instructor characteristics. Had the effects been of significance the results would not be used to adjust the scores since rank and sex of the instructor can not be
considered exogenous variables. Discovery of significant bias for these variables could, however, lead to further studies designed to examine the causes of such bias.

Arguably, it may be that the rank and sex of the instructor play a part in the assignment of courses (particularly to lower level courses) even though the effects of rank and sex of the instructor alone are not significant. Efforts to model rank on course level variables were not significant. However, Model 2 in Table 5 shows the effect of the sex of the instructor on the intercept and course levels. The coefficient for female instructors modeled on the intercept (in this case male instructors teaching graduate courses), was again insignificant. However, modeling sex on course level produced significant results, which indicates that the effect of sex has been moderated by the effects of course level. This is an interaction effect. For example, a male instructor teaching a graduate course would have a lower estimated score (5.64) than a female instructor (5.71) but this difference would not be significant. However, a male instructor teaching a LEV123 course would have an estimated score of 5.54 (5.636 -.101) compared to an estimated score for a female instructor of 5.42 [5.636 + .077 -.101+.191]). This difference represents 27% of a standard deviation.

Similarly a male instructor would have an estimated score for a LEV4 course of 5.56 (5.636 -.072) compared to an estimated score for a female instructor of 5.46 [5.636 + .077 -.072 + .185]). This represents 25% of a standard deviation.

As was previously discussed, the practice of adjusting scores on the basis of sex of the instructor is problematic, particularly when the reason for this bias is unknown. This model provides additional information as to why there are differences in the scores for male and female instructors. If female instructors are habitually assigned to teach courses in which the
estimated scores are lower because of other factors (for example, percentage of females) the cumulative effect of these variables plus sex may lead to substantive differences in the scores. However, without direct knowledge as to why female instructors are rated lower in undergraduate courses, efforts to adjust their scores might be perceived as unfair. The information from this model may not be used for score adjustment, but may cause administrators to more closely examine patterns of teaching assignments within the faculty.

The percentage of variance explained is indicated at the bottom of the table. Despite the inclusion of level-2 predictors, there was no substantial increase in the amount of variation explained. However in both these models the reliability remains quite high at .79.

4.4 Longitudinal Models

The final stage of the analysis considered whether the variance within- and between-instructors could be explained by course-level variables representing time. The previous models served to explain a substantial portion of the variance based on average scores for each instructor. Because summative evaluations of faculty usually contain scores for an instructor over a period of time, the model was extended to include variables which served as measures of change.

The first analysis used a two-level model to examine the effect of time, where time was the date of the evaluation. The results are presented in Table 6. The findings show that both time and its square were significant, indicating a nonlinear relationship between time and the scores. As with class size, these coefficients are treated as a unit in order to examine
the polynomial relationship. Fig 3 demonstrates the estimated trend in the scores. The trend is positive with a modest negative acceleration of growth over time.

The variance of the intercept is significant, meaning that we can distinguish between instructors in their initial status in the model. In this model, however, this statistic is not particularly meaningful because it includes the projection of faculty growth scores back to 1982. All of the instructors in the sample did not teach in 1982, so the regression estimates the intercept for each of these instructors based on the information obtained from faculty who had taught at that time. Large numbers of such estimates contribute to a lower estimate of the reliability of the intercept (.63).

![Fig. 3 Predicted SET Score Over Time](image)

The variances of time and its square were also significant, indicating that we can distinguish between instructors for their individual growth trajectories. The reliability for
### Table 6
HLM Analysis of the Effects of Course-Level Variables on Course Means
Within Instructors: Time Model

<table>
<thead>
<tr>
<th>Composite Score</th>
<th>Coefficient</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Within Instructor Equation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>5.671**</td>
<td>0.034</td>
</tr>
<tr>
<td>TIME</td>
<td>0.024**</td>
<td>0.005</td>
</tr>
<tr>
<td>TIMESQD</td>
<td>-0.001**</td>
<td>0.001</td>
</tr>
<tr>
<td>PCTELEC</td>
<td>0.003**</td>
<td>0.001</td>
</tr>
<tr>
<td>PCTFEM</td>
<td>0.001*</td>
<td>0.001</td>
</tr>
<tr>
<td>LEV 123</td>
<td>-0.117*</td>
<td>0.048</td>
</tr>
<tr>
<td>LEV4</td>
<td>-0.080</td>
<td>0.046</td>
</tr>
<tr>
<td>SIZE</td>
<td>-1.289**</td>
<td>0.148</td>
</tr>
<tr>
<td>SIZESQD</td>
<td>6.466**</td>
<td>0.997</td>
</tr>
<tr>
<td>SIZECBD</td>
<td>-6.930**</td>
<td>1.373</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Random Effects</strong></th>
<th>SD</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.419</td>
<td>0.176**</td>
</tr>
<tr>
<td>TIME</td>
<td>0.045</td>
<td>0.002**</td>
</tr>
<tr>
<td>TIMESQD</td>
<td>0.006</td>
<td>0.000**</td>
</tr>
<tr>
<td>PCTELEC</td>
<td>0.004</td>
<td>0.000**</td>
</tr>
<tr>
<td>LEV 123</td>
<td>0.417</td>
<td>0.174**</td>
</tr>
<tr>
<td>LEV4</td>
<td>0.390</td>
<td>0.153**</td>
</tr>
<tr>
<td>level-1</td>
<td>0.498</td>
<td>0.248</td>
</tr>
</tbody>
</table>

**Model Statistics**

<table>
<thead>
<tr>
<th>Reliability of Estimates</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.627</td>
</tr>
<tr>
<td>TIME</td>
<td>0.399</td>
</tr>
<tr>
<td>TIMESQD</td>
<td>0.132</td>
</tr>
<tr>
<td>PCTELEC</td>
<td>0.356</td>
</tr>
<tr>
<td>LEV123</td>
<td>0.417</td>
</tr>
<tr>
<td>LEV4</td>
<td>0.414</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percent Variance Explained</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Within Instructor</td>
<td>25</td>
</tr>
<tr>
<td>Between Instructor</td>
<td>20</td>
</tr>
</tbody>
</table>

**p < .01  * p < .05**
these variables was quite low which may be due in part to the fact that the time variables were random, and therefore trends for individual faculty members would vary in terms of their slopes. However, the low reliability of the estimates for time and its square may not indicate a lack of precision, since if all teachers improved at a similar rate, this would cause the reliability of the scores to be low (see Rogosa & Willett, 1983; Willett, 1988). As with the intercept, the trends for faculty who had not taught over the entire time period would be estimated, also contributing some unreliability to the model.

The effects of percentage of students taking the course as an elective, percentage of females, course level and class size remained significant, even after controlling for the effects of time. The percentages of variance explained by the course-level model in Table 4 were 19.8% and 16.6% for within instructor and between instructor respectively. In this model, those figures have increased to 25% and 20%, which shows that time explains an additional 5.2% and 3.4% percent of the variance. The model shows that over time, on average, we would expect scores to improve. What we do not know is whether the increase is due to increased experience, increased age of the instructors or some other variable such as policy changes or changes in students’ perceptions of instruction.

The model has utility as a means of tracking the overall trends in student ratings over time. It provides a tool for administrators to demonstrate continuing quality of teaching and demonstrates the overall effects of faculty development programs. However, it may not be the best model for adjusting individual scores since the selection of the time period is arbitrary, and fails to take into account the experience of the individual instructor. It would not be fair to assume that two instructors teaching in 1986 should have similar scores if one
instructor had twenty years of experience and the other was teaching for the first time. Also, since the scores increase gradually over time the score adjustment would be made towards the baseline level in 1982, a nonproductive practice if you want to show improvement in the scores.

Table 7 shows the results for the longitudinal model based on the years of experience of the instructor who taught the course. Period one (PER1) represented the first five years of the instructors’ career, Period 2 (PER2) represented years 5 to 10 in the career path, Period three (PER3) represented years 10 to 15, Period four (PER4) represented years 15 to 25, and period 5 (PER5) the years 25 to 35.

With the exception of Period 4, 15 to 25 years of experience, all of the coefficients were significant. There was a significant increase in scores in the first 5 years, from 10 to 15 years and from 25 to 35 years. This trend is presented in Figure 4. To obtain the scores used to plot this trend the years of experience of the instructor at the time the course was taught was multiplied by the period in which the course was taught and then multiplied by the coefficient for that period. This value was added to the value for the intercept. For example, a course taught in the fifth year of an instructors’ career would have an estimated score of 5.82 \[5.552 + (5)(.053)\]. A course taught in the tenth year of an instructors’ career would have an estimated score of 5.16 \[5.552 + (10)(-.039)\].

For 5 to 10 years of experience, there was a decrease in the scores, but this coefficient was also random. The significance of the variance of this random effect indicates that we can distinguish between the instructors for this time period. This period of time in the career paths of faculty may be of particular interest to those in faculty development in terms of
tracking which faculty improve in their teaching, and why, and also targeting those in need of training. It would also be of interest to note which faculty members received promotion at the end of the five year period and how this relates to the scores from their teaching evaluations.

![Fig. 4 Predicted SET Score by Years of Experience](image)

The reliability for this model was also lower than the previous models for similar reasons similar to the time model. The sample would not have included data from the first five years of teaching for faculty who began teaching earlier than 1982, and therefore these estimates would have been estimated with less precision. The sample would not have included data for all instructors for all time periods which may also have contributed to lower reliability for this model.

This model drastically changed with regard to the percent variance explained. Since the amount of within-instructor variance exceeded that of the null model new percentages
were computed. The results showed that the amount of variance explained within-instructors was 52%. The variance explained between-instructors was 48%. This is a substantially larger amount of explained variance than has been previously reported in the literature (approximately 16%, see Marsh, 1987).

As with sex and rank of the instructor, it is difficult to make the case that the experience of the instructor is an exogenous variable that biases the scores. It could be argued that obtaining experience is simply an artifact of staying on faculty—all instructors will gain experience if they remain employed by the university. Thus the use of this information to adjust scores may not be appropriate. Because this model explains a large amount of the variance both within- and between-instructors it has potential for both interpretation of an individual instructors’ scores over time and for comparisons of faculty used for awards and recognition. The information that on average there is some fluctuation in instructors scores over time coupled with the substantial increase in within-instructor variation explained by the model indicates that administrators need not be concerned with some fluctuation in the scores of an instructor over time. However, a consistent decrease in scores, particularly after the tenth year of experience, would be cause for concern since that pattern would deviate from the prevailing trend in the faculty which shows that scores increase with experience.

The substantive influence of experience on the variation between instructors demonstrates that experience has some influence on the scores and that comparisons between faculty who are in different periods of their careers may not be appropriate. While score adjustment based on these data may not be used, these results could be used to establish separate criterion for each of the time periods.
Table 7
HLM Analysis of the Effects of Course-Level Variables on Course Means
Within Instructors: Instructor Experience Model

<table>
<thead>
<tr>
<th></th>
<th>Composite Score</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed Effects</strong></td>
<td><strong>Coefficient</strong></td>
<td><strong>SE</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Average Within Instructor</strong> Equation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>5.552**</td>
<td>0.067</td>
<td></td>
</tr>
<tr>
<td>PCTELEC</td>
<td>0.003**</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>PCTFEM</td>
<td>0.001**</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>LEV 123</td>
<td>-0.146**</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>LEV4</td>
<td>-0.111*</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td>SIZE</td>
<td>-1.219**</td>
<td>0.149</td>
<td></td>
</tr>
<tr>
<td>SIZESQD</td>
<td>6.455**</td>
<td>1.005</td>
<td></td>
</tr>
<tr>
<td>SIZECBD</td>
<td>-7.019**</td>
<td>1.384</td>
<td></td>
</tr>
<tr>
<td>PER1</td>
<td>0.053**</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td>PER2</td>
<td>-0.039**</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>PER3</td>
<td>0.023**</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>PER4</td>
<td>0.003</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>PER5</td>
<td>0.026*</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td><strong>Random Effects</strong></td>
<td><strong>SD</strong></td>
<td><strong>Variance</strong></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.530</td>
<td>0.176**</td>
<td></td>
</tr>
<tr>
<td>PCTELEC</td>
<td>0.004</td>
<td>0.000**</td>
<td></td>
</tr>
<tr>
<td>LEV 123</td>
<td>0.434</td>
<td>0.174**</td>
<td></td>
</tr>
<tr>
<td>LEV4</td>
<td>0.423</td>
<td>0.153**</td>
<td></td>
</tr>
<tr>
<td>PER2</td>
<td>0.104</td>
<td>0.011**</td>
<td></td>
</tr>
<tr>
<td>level-1</td>
<td>0.509</td>
<td>0.259</td>
<td></td>
</tr>
</tbody>
</table>

Model Statistics

Reliability of Estimates

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.516</td>
</tr>
<tr>
<td>PCTELEC</td>
<td>0.374</td>
</tr>
<tr>
<td>LEV123</td>
<td>0.479</td>
</tr>
<tr>
<td>LEV4</td>
<td>0.496</td>
</tr>
<tr>
<td>PER2</td>
<td>0.368</td>
</tr>
</tbody>
</table>

Percent Variance Explained

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Within Instructor</td>
<td>52</td>
</tr>
<tr>
<td>Between Instructor</td>
<td>48</td>
</tr>
</tbody>
</table>

** p < .01  * p < .05
Summary

This analysis has employed HLM and three types of models to examine student ratings of instruction: the null model and course- and instructor-level models. Further discussion of these results occurs in the final chapter, which includes a discussion of the utility of this research and implications for further study.

Results of the null model showed that there was considerable variation to be explained both within- and between-instructors, with more variation occurring within-instructors. The null model was also used to assess the potential use of the Composite Score and the Global Item. The mean and variability for both scores were quite similar and both models provided reliable estimates: therefore either score could be used with confidence. The analysis also showed that the minimum number of courses required per instructor to provide a reliable estimate of teaching effectiveness was 5.

Course-level models were used to assess the effects of exogenous variables on the scores, and to examine two types of longitudinal models. The percentage of students taking the course as an elective, the percentage of females taking the course, course level and class size all had a significant effect on the scores, and explained a substantial proportion of both the within- and between-instructor variance. Results of the instructor-level models showed that the effects of sex and rank of the instructor where not significant or conclusive.

Results of the first longitudinal model were significant, indicating that it is possible to distinguish between instructors for their individual growth trajectories. The second longitudinal model demonstrated that the experience of the instructor explained a substantive amount of both the within- and between-instructor variance.
Chapter Five: Summary

Student evaluations of instruction provide data that can be used to make administrative decisions about promotion, tenure, awards and merit increases for tenure-track and tenured faculty. Despite considerable research, questions about methodology and interpretation of the results persist. The purpose of this study was to specify and test an Hierarchical Linear Model to address research questions concerning the reliability of the scores, adjustment for bias and long-term stability. The study employed ratings of faculty who had been evaluated over an eleven-year period to investigate these questions. This chapter discusses the policy implications of the results.

Several analyses were performed using HLM, a statistical technique appropriate for the analysis of hierarchical or nested data. It is used here to perform multilevel regression analyses on data using two units of analysis, or levels: the course and the instructor. These data can be seen to be hierarchical, as the courses are nested within the instructors who teach them. This type of analysis allows the variation in the scores on student evaluations to be partitioned into within-instructor and between-instructor components, and allows the relationship of scores and variables from both the within-instructor and between-instructor levels to be examined.

First, the null model was estimated to determine the most appropriate score to be used for summative evaluation and the number of courses per instructor required in order to obtain a reliable estimate of an instructors' teaching effectiveness. Course- and instructor-level models were estimated to examine the effects of exogenous and instructor variables on the
scores. Two longitudinal models were designed to examine the effects of time and experience of the instructor on the scores.

The study differed from previous research in several ways. First, most previous research on stability of student evaluations of instruction used two scores for each instructor which prevented the researchers from examining the long-term stability of the scores. The few examples of longitudinal research did not include the examination of course and instructor variables in conjunction with time, limiting the utility of that research for adjustment of the scores. Third, the use of HLM provides the opportunity to partition the variance into within- and between-instructor variation. Designs in which the courses for each instructor are pooled mask the within-instructor variance: variation in the scores is assumed to be either between-instructor variance or measurement and sampling error. The nature of this study allows for the simultaneous examination of within- and between-instructor variance. It specifies a multivariate model for assessing potential sources of bias and examining changes in scores over time.

The analysis of the null models suggests that the Composite Score and the Global Item might be used interchangeably for making personnel decisions, since the means and standard deviations, the amount of variation partitioned within- and between-instructors and the reliabilities are quite similar. However, other reliability concerns related to class size and the number of items make the Composite Score a better choice. In the context of a faculty in which there is considerable variability both in the number of occasions on which each instructor has been evaluated and the size of classes performing the evaluation, it is more appropriate to use the 30-item mean in order to bolster the reliability of the estimates.
The selection of an appropriate level of reliability should not be arbitrary or driven solely by theory. The nature of the decisions to be made plays an important role in setting this level as does the total workload for each instructor. The number of courses evaluated for each instructor was quite variable and the total for an instructor was often too small to set a standard for reliability higher than .70 to .80. However other duties that comprise the workload of an individual, such as research and administrative duties, must also be taken into consideration when summative evaluations are made.

The findings from the course level model are consistent with Centra’s (1993) argument, in which he stipulated that when several factors are taken together in a multivariate analysis, the sum of their effects on the score may lead to significant results. Results of the analysis show that the percentage of students taking the course as an elective, the percentage of females in the class, class size and the level of the course all had a significant effect on the scores. The Hierarchical Linear Model generated coefficients that can be used for adjustment of the scores. Taken together, they represent a substantial effect on the outcome of the scores, and adjustment for these factors could have a substantive impact on the final score an instructor receives.

The worst case scenario would be the cumulative effect of these variables on a course in which all the students were enrolled as a requirement, in which there were no females, the course was a 100, 200 or 300 level course and the class size was 98 (the largest in the sample). Using the coefficients from the model, the estimated average score for a course with the above-mentioned characteristics would be 4.59 [i.e. $5.656 + .003(0-34.19) + .001(0-67.35) - .159 + (98-17.39/100)(-1.5) + (98-17.39/100^2)(6.522) + (98-17.39/100^3)(-7.196)$].
The estimated score for an average course was 5.66. The difference between the estimate for an average class and this hypothetical class represents 2.5 standard deviations—a substantial difference. Adjustment of the scores to account for these characteristics would benefit this instructor by placing his or her score in a more acceptable range.

This example may be somewhat overstated, given the nature of this sample with respect to class size. Few courses in the sample are this large, and it may be that another variable, such as the type of course or course level is confounding the results for large classes. Until further analysis is carried out to investigate this possibility, score adjustment for large class sizes should be used with caution for this faculty.

Results of the analysis involving instructor characteristics showed than rank and sex of the instructor did not have a significant effect on the scores. This is an important finding, as it demonstrates to administrators that the evaluations are not biased by these factors, and assures faculty members that their scores are not significantly affected by these factors over which they have no control. When sex of the instructor was modeled on the level of the course, the results were significant. The estimated scores for females instructing undergraduate courses would be lower than the estimated scores for males. Despite the significant results the scores would not be adjusted for sex of the instructor, since this variable can not be considered exogenous. However these results do provide important information on which to base further examination of course assignments within the faculty.

When scores of individual faculty are to be compared, it is appropriate to take the experience of the teacher into consideration. The first longitudinal model used chronological time as the longitudinal variable. Results indicated that there was a significant but slight
increase in scores over time, but these results only reflected the faculty trend. The model provides administrators with a general picture which demonstrates the overall quality and consistency of instruction over time and shows that student ratings of instruction provide the basis for a continuing record of progress if they are continuously administered. It does not provide a model on which to compare faculty since for any given time in the data, different faculty members have various levels of teaching experience and familiarity with the particular courses they are teaching.

The second longitudinal model provides an estimate of the average performance of the instructors over time, when time is defined as the period of time in the career path of a tenure-track instructor. On average, instructors in the first five years, ten to fifteen years, fifteen to twenty-five years and twenty-five to thirty years showed increases in the scores that were fixed. Results for the period from five to ten years of experience showed post-tenure decline and were random, meaning that the scores were quite variable in this time period. These results provide administrators with knowledge as to when scores should be closely monitored to determine if faculty members require inservice, and may lead to further investigation as to the relationship of scores from student evaluations of teaching and career advancement. The increase in scores over the first five years of the candidate’s career also indicates to administrators that on average, scores for faculty members are likely to be at their lowest during the first two years. Since reappointment decisions are often made after two years, knowledge of this particular trend may be crucial for untenured faculty.

Only a small number of the potential course and instructor characteristics have been used for these models, with good results. The variables in the analysis represented a
significant contribution to the variance, but additional research is required to better specify the model and to assess the effects of other potentially influential factors. For this particular institution factors relating to mature students and international students may be important variables to study. However, future developments in research should not preclude the use of these coefficients for score adjustments based on the course-level model, since the reliability of the model was quite high and the effects of the course-level variables remained significant even after controlling for the longitudinal variables and the instructor characteristics.

The advantage of this data set included the large number of instructors involved, the size of the within-instructor samples and the amount of longitudinal information on individuals. This analysis aggregated student-level variables to the course-level and because the evaluators remain anonymous, the courses attended and evaluated by individual students can not be used as a variable. Given that students may vary drastically on their ratings due to some personal variable, a more powerful test could be achieved with a three-level analysis (students, courses, instructors). This study has limited generalizability in that the sample is derived from a single faculty. An Hierarchical Linear Model including larger units, such as academic units, departments or perhaps universities would be an important contribution to research on context, particularly if the evaluation instrument and procedures are duplicated across each of the settings.

Further research is also required with regard to policy issues concerning the setting of appropriate norms or criteria for faculty evaluation. HLM provides a model that can be used to adjust scores, but must be used in the framework of standards deemed fair for instructors
in a particular institution. With the appropriate standards in place, HLM becomes a tool that can help ensure that instructors are provided equitable results.
References


STUDENT BACKGROUND INFORMATION

Date of responding to this questionnaire:

Month _______ Year _______

Your sex:

☐ Male ☐ Female

Your age:

☐ 24 or less ☐ 25-29 ☐ 30-34
☐ 35-39 ☐ 40-44 ☐ 45+

In your case this course is:

☐ A requirement ☐ An elective

Please indicate your program of studies.
(please fill only one bubble of the six below)

Teacher Education Program:

☐ Elementary Teacher Certification (12 month option)
☐ Elementary Teacher Certification (2 year option)
☐ Secondary Teacher Certification

or

☐ Diploma Program
☐ Masters or Doctoral Program
☐ Unclassified Student

OVERALL RATING

In the Faculty of Education, students are asked to evaluate the quality of instruction in every course. In responding to the items given here, please keep in mind that you are asked to rate the instructor and those aspects of the course over which he or she clearly has control. Please do not evaluate other related factors, such as department or faculty policies. This is to be an evaluation of the individual instructor's teaching and related activities and outcomes.

Considering all facets of this instructor's teaching, I would rate it as:

(fill only one of the bubbles below)

☐ Excellent
☐ Very Good
☐ Good
☐ Adequate
☐ Less than adequate
☐ Poor
☐ Very Poor

Please complete the back page which has 30 items pertaining to specific aspects of the quality of instruction in this course.
### Evaluation Form

For the remaining items, a seven-point rating scale is provided for you to record your responses. Please interpret it to have the following meaning:

<table>
<thead>
<tr>
<th>Rating</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>1</td>
<td>Disagree Very Strongly</td>
</tr>
<tr>
<td>2</td>
<td>Disagree Somewhat</td>
</tr>
<tr>
<td>3</td>
<td>Agree Somewhat</td>
</tr>
<tr>
<td>4</td>
<td>Agree</td>
</tr>
<tr>
<td>5</td>
<td>Agree Very Strongly</td>
</tr>
</tbody>
</table>

Please keep in mind that a rating of '7' is ALWAYS intended to reflect an extremely FAVOURABLE reaction to instruction in your course, and a rating of '1' is ALWAYS intended to reflect an extremely UNFAVOURABLE reaction.

If you feel that any item is totally irrelevant to the quality of instruction in your course, completely fill the '0' (zero) bubble provided. This may occur, for example, in the case of an item which asks about exams in a course which has no exams.

1. The amount of work required by the instructor in this course is reasonable. .............................................. 0 1 2 3 4 5 6 7
2. The instructor's class presentations are well organized. ................................................................. 0 1 2 3 4 5 6 7
3. Assignments required by the instructor are useful learning experiences. ............................................. 0 1 2 3 4 5 6 7
4. The instructor is teaching me to learn more on my own in this subject. .............................................. 0 1 2 3 4 5 6 7
5. The instructor uses scheduled class time efficiently. ................................................................. 0 1 2 3 4 5 6 7
6. When students are confused, the instructor attempts to clarify. .......................................................... 0 1 2 3 4 5 6 7
7. Lectures are consistent with the subject matter in the course outline. .............................................. 0 1 2 3 4 5 6 7
8. The instructor has helped me to understand relationships among important topics and ideas. .................. 0 1 2 3 4 5 6 7
9. The level of intellectual activity required by the instructor is appropriate for this type of course. ............. 0 1 2 3 4 5 6 7
10. The instructor communicates clearly in class. ......................................................................................... 0 1 2 3 4 5 6 7
11. Evaluation procedures have allowed me to demonstrate adequately what I have learned. ........................ 0 1 2 3 4 5 6 7
12. The teaching in this course has motivated me to study topics on my own initiative. ............................. 0 1 2 3 4 5 6 7
13. The instructor provides adequate class time for questions. ............................................................... 0 1 2 3 4 5 6 7
14. The instructor presents concepts in a manner that aids student learning. ............................................. 0 1 2 3 4 5 6 7
15. Evaluation procedures (exams, grading, etc.) are fair. ................................................................. 0 1 2 3 4 5 6 7
16. As a result of taking this course, I have become more interested in the subject. ..................................... 0 1 2 3 4 5 6 7
17. The instructor emphasizes outcomes which most students can achieve in the time allocated. .................. 0 1 2 3 4 5 6 7
18. The instructor is available to help students outside of class. ............................................................ 0 1 2 3 4 5 6 7
19. The text (if selected by the instructor) and other reading materials have been important to my understanding of the course. ......................................................... 0 1 2 3 4 5 6 7
20. The instructor has helped me to do work that is the highest quality I can achieve. .......................... 0 1 2 3 4 5 6 7
21. Course difficulty established by the instructor seems appropriate for most students. .............................. 0 1 2 3 4 5 6 7
22. The instructor's responses to questions are clear. .................................................................................... 0 1 2 3 4 5 6 7
23. Assignments contribute to the achievement of course objectives. .......................................................... 0 1 2 3 4 5 6 7
24. My knowledge and skills have increased as a result of taking this course. ................................................ 0 1 2 3 4 5 6 7
25. The instructor has made course objectives and requirements clear. .................................................... 0 1 2 3 4 5 6 7
26. The instructor seems to take account of students' abilities, needs, and interests. ..................................... 0 1 2 3 4 5 6 7
27. Evaluation procedures provide useful feedback for student improvement .................................................. 0 1 2 3 4 5 6 7
28. I would recommend this instructor's course to other students wishing to acquire knowledge and skills in this area. .......................................................................................... 0 1 2 3 4 5 6 7
29. The instructor stated clearly the basis for evaluating students. ............................................................ 0 1 2 3 4 5 6 7
30. The basis for evaluation, as stated, was followed. ................................................................................... 0 1 2 3 4 5 6 7