2010

Student Perception Of General Education Program Courses

Julie Pepe
University of Central Florida

STUDENT PERCEPTION OF GENERAL EDUCATION PROGRAM COURSES

by

JULIE WILDMAN PEPE
B.A. Rollins College, 1982
M.S. Purdue University, 1984

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in the College of Education
at the University of Central Florida
Orlando, Florida

Summer Term
2010

Major Professors: Lea Witta and Morgan Wang

ABSTRACT

The purposes of this study were to: (a) determine, for General Education Program (GEP) courses, what individual items on the student form are predictive of the overall instructor rating value; (b) investigate the relationship of instructional mode, class size, GEP foundational area, and GEP theme with the overall instructor rating value; (c) examine what teacher/course qualities are related to a high (Excellent) overall evaluation or a low (Poor) overall evaluation value.

The data set used for analysis contained sixteen student response scores (Q1-Q16), response number, class size, term, foundational area (communication, cultural/historical, mathematics, social, or science), GEP theme (yes/no), instructional mode (face-to-face or other), and percent responding (calculated value). All identifying information such as department, course, section, and instructor was removed from the analysis file. The final data set contained 23 variables, 8,065 course sections, and 294,692 student responses.

All individual items on the student evaluation form were related to the overall evaluation item score, measured using Spearman's correlation coefficients. None of the examined course variables were selected as significant when the individual form items were included in the modeling process. This indicated students employed a consistent approach to the evaluation process regardless of large or small classes, face-to-face or other instructional modes, foundational area, or percent responding differences.

Data mining modeling techniques were used to understand the relationship of individual item responses and additional course information variables to the overall score. Items one to fifteen (Q1 to Q15), class size, instructional mode, foundational area, and GEP theme were the independent variables used to find splits to create homogenous groups in relation to the overall evaluation score. The model results are presented in terms of if-then rules for “Excellent” or “Poor” overall evaluation scores. The top three rules for “Excellent” or “Poor” based their classifications on some combination of the following items: *communication of ideas and information; facilitation of learning; respect and concern for students; instructor’s overall organization of the course; instructor’s interest in your learning; instructor’s assessment of your progress in the course; and stimulation of interest in the course.* Proportion of student responses conforming to the top three rules for “Excellent” or “Poor” overall evaluation ranged from 0.89 to .60.

These findings suggest that students reward, with higher evaluation scores, instructors who they perceive as organized and strive to clearly communicate course content. These characteristics can be improved through mentoring or professional development workshops for instructors. Additionally, instructors of GEP courses need to be informed that students connect respect and concern and having an interest in student learning with the overall score they give the instructor.

This dissertation is dedicated to my parents, Jane and George Wildman. Thank you for your guidance, support and love! You taught me so many important lessons about life; the wonder of knowledge discovery is just one example.

ACKNOWLEDGMENTS

I would like to express my appreciation to my dissertation co-chairs, Dr. Lea Witta and Dr. Morgan Wang for their support and guidance. To the other members of my committee, Dr. Charles Dziuban, Dr. Debbie Hahs-Vaughn, and Dr. Mark Johnson, my sincere thanks for spending your valuable time reviewing my work and offering helpful suggestions.

A special thanks to Dr. Robinson, Dr. Boote, and Dr. Hayes for making my degree program a reality. You made the possibility of combining my data mining training with the PhD curriculum happen.

Sincere thanks to my friends and colleagues (Anne Grey, Ana Leon, Tace Crouse, Alison Morrison-Shetlar, Corine Strebel, Susan Schott, Connie Cutchins, Don Worchester, and Elena Sequera) that have provided encouragement throughout this entire process. To my confidant, Joanne Roche, who passed away last year, I think of you often and miss your smiling face greeting me every work day. God bless you.

Finally and most importantly, a huge heartfelt thanks must go to my husband, David and my children, Ben and Mark for keeping me going during this six year process. Without your support and love my journey would be incomplete.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

The phrase “the consumer is always right” has been a motto for success, but does the practicality of this phrase transcend the business world and apply to the academic arena? In higher education a current focus on the educational consumer is one driver of the accountability movement that pervades American learning culture. Because of the amount of money being spent on education this trend is gaining momentum (Gravestock & Gregor-Greenleaf, 2008; Landrum & Braitman, 2008). This emphasis is further supported by accreditation requirements for assessing student learning (Association of American Colleges and Universities, 2008; Council for Higher Education Accreditation, 2008). Seldin (1999) reported that, “student ratings are now the most widely used source of information on effective teaching” (p. 15). In 2008, the federal government provided $52.1 billion in financial aid support to students attending a college or university (National Center for Education Statistics, 2009). Since there are many options available to students today, it is particularly important that higher learning institutions understand how students view their educational experience.

This Chapter will describe important aspects of student evaluation as it pertains to this study. Identification of the study purpose, research questions, definition of terms, and significance of the study will follow the theoretical background. A brief description of the study methodology, limitations and assumptions of the study, conclude the introduction.

Theoretical Background

There is no denying the importance of the student’s role in evaluating instruction in higher education. D’Apollonia and Abrami (1997) reported that 98% of higher education institutions in the United States use some form of student evaluation and that an increasing percentage of international institutions are doing so as well (Moore & Kuol, 2005). The fact is, evaluation information is being used to make policy and personnel decisions and will continue to be utilized by students, department heads, and administration for an even broader array of purposes (Kulik, 2001). Abrami, Theall, and Mets (2001) reported that student information "serve as tools for instructional improvement, as evidence for promotion and tenure decisions, as the means for student course selection, as one criterion of program effectiveness, and as the continuing focus of active research and intensive debate" (p.1).

In the late 1920's publications on student evaluation results began appearing in journals. Most of the earlier research was focused on issues surrounding the specific construction of the evaluation form. Validity and reliability studies dominated the research in the 1980's. In the 1990's studies regarding student, instructor, or classroom conditions that might have an influence on student scores were prominent. Abrami, d'Apollonia, Centra, Feldman, Marsh, Roche, and Seldin are commonly cited authors on the topic of student evaluation. In the past decade, the potential relationship between scores and grades or expected grades has received the most attention.

The primary source of student based faculty evaluation comes in the form of an institutionally standardized questionnaire distributed to students in the final weeks of a course (Abrami, 2001). Information from these questionnaire protocols is usually the most complete data set available to assess student views on the course and the instructor (Moore & Kuol, 2005). The terms used to identify student evaluation forms are typically known by the following acronyms (Abrami, Theall, & Mets, 2001):

SET – Student Evaluation of Teaching

SRTE - Student Ratings of Teaching Effectiveness

TCE – Teacher Course Evaluations

TRF – Teacher Ratings Forms

The components or items included in the evaluation form vary by institution, thus making direct comparison difficult. Questionnaire construction may focus on different aspects of instruction, thus yield very different results. Criticism of student evaluation information is directed at how the information is used, what the values indicate, and what overall conclusions can be reached (McKeachie, 1997). In general, student ratings appear to have two main functions: formative and summative. Formative evaluation provides faculty with information to improve aspects of instruction. Preferably formative evaluation information is kept confidential and not used in personnel or program decisions (Theall & Franklin, 2001). Additionally some mentoring system should be put in place to assist and support faculty throughout the evaluation process.

Summative information is used to decide tenure and promotion status, teaching awards, internal reporting, raises, or termination. The deleterious aspect of using student evaluation scores comes when this information is the only data used for assessing quality instruction. Cashin (1999) reported, “Many colleges and universities rely heavily, if not solely, on student rating data as the only systematic source of data collected to evaluate teaching” (p. 26).

Empirical studies use class, student, and teacher characteristics as a gateway to understanding what factors are identified as statistically significant (i.e., important) in the numerical scores collected from student evaluation forms. As for the meaning of the evaluation scores, some researchers agree they measure several aspects of effective teaching while others believe they measure student satisfaction (Abrami, d’Apollonia, & Cohen, 1990; Beyers, 2008; Centra, 1993; Marsh & Roche, 1997). A generally accepted definition of effective teaching has not been determined (Trout, 2000; Paulsen, 2002). McKeachie (1997) notes the meaning of effective teaching has not been defined and depends on the goals for instruction. Kolitch and Dean (1999) say that an effective teacher will be able to communicate clearly, be organized, and interact well with students via examples and relevant questions.

This study focuses on the use of student evaluation information to examine their perceptions of General Education Program (GEP) courses at a large public institution. GEP courses are designed to introduce students to a wide-range of areas they may not otherwise be exposed to in their respective major. These required courses comprise what is known as general education curriculum or core curriculum, depending on the institution (Sudermann, 1992). Although, research on student evaluation information has been extensive, empirical studies for GEP courses are non-existent.

Data mining techniques are a relatively new collection of statistical methods that apply to analyzing very large data sets to maximize extraction of information (Hand, Mannila, & Smyth, 2001). Data mining methodology and associated tools, such as decision tree analysis, allows all responses to be utilized, which in this dissertation comprises several hundred thousand observations. Data mining tools have no strict assumptions for the functional form of the model, are robust against the presence of outliers, and are resistant to the curse of dimensionality (Wang, 2007). Decision tree analysis is a flexible modeling tool that is an efficient method for studying large data sets (Wang, 2007). Missing values, which may be substantial, do not need to be imputed. Software options allow missing data values to be included as a separate category in initial modeling stages and surrogate rules can be implemented for additional data set modeling results (SAS User's Manual, 2009).

The goal of this study is to find questionnaire items that are related to the overall score provided by students and then use this relationship to identify decision rules for predicting the overall rating value. The data used for initial model construction and validation are stratified samples of the entire data set. The first modeling stage uses a percentage of the data to construct the model; the model is then refined using the remaining data. This procedure allows for a true measure of predictive model validity to be determined (Breiman, Friedman, Olshen & Stone, 1984).

Purpose

The purpose of this study was to determine, for GEP courses, what instructor qualities can be identified as critical factors for scores obtained via student evaluation forms. These factors were compiled using the individual items on the student evaluation form or course factors that are predictive of the overall instructor rating value. The study examined what instructor/course qualities are related to a high overall evaluation or a low overall evaluation value thus placing an overall evaluation in the context of actionable items for improvement. Additional information on course year, instructional mode, GEP foundational area, and class size was examined in relation to overall scores given by students. This research contributes unique information to the topic of instructor evaluations by specifically focusing on GEP courses and the influence of program construction (i.e., class size, theme, foundational area) on student perception.

Research Questions

1. What items on the student evaluation form are related to the “Overall rating” (item 16) score provided by students?

2. Is instructional mode, class size, GEP foundational area, or incorporating a GEP theme related to the "Overall rating" (item 16) score provided by students?
3. What rules can be identified to understand the determination of an "Excellent" overall rating on the student evaluation form?
4. What rules can be identified to understand the determination of a "Poor" overall rating on the student evaluation form?

Definition of terms

Foundational area – Categorical variable representing the foundational area of the course as defined in the institution's course catalog.

General education program (GEP) – Collection of courses that provide students with a common foundation of study. This study uses data collected from courses listed in the university course catalog under general education program requirements. A summary of these courses (by year) is provided in Appendix B.

GEP theme –Overarching topic meant to connect student understanding of a central issue via course content. Global climate change has been the common theme for the academic years 2006-2007, 2007-2008, and 2008-2009.

Individual items – Set of 15 items (Q1 to Q15) on the student perception of teaching form (Appendix A).

Overall rating (Q16) – Final item on the student evaluation form which reads, "Overall assessment of instructor." The responses are on a five-point Likert scale from 1 to 5.

Percent responding – Calculated value using number of student responses divided by number of enrolled students.

Student evaluation form –Institution specific questionnaire with a set of 16 items (on a 1 to 5 scale), as given in Appendix A.

Significance of the Research

Colleges and universities today are under increasing pressure to show evidence of quality instruction (Joe, Harmes & Barry, 2008). Student perception information is typically a very complete data set generated from a systematic approach and using this information as part of the bigger picture of teaching effectiveness is a generally accepted practice (Algozzine et al., 2004; Marsh, 1984; Stark-Wroblewski, Ahlering, & Brill, 2007; Theall & Franklin, 2001).

Wang, Dziuban, Cook, and Moskal (2009) published results generated from a similar data set containing all courses from a large public institution for academic years 1996-2001. Their study used data mining techniques to develop a model for overall evaluation rating using the 16 individual SPI items, course level (lower undergraduate, upper undergraduate or graduate) and academic year (1996-97, 1997-98, or 1998-99). The decision tree results focused on three qualities of effective teaching: effective communication skills, facilitative teaching, and organization and assessment (Wang et al., 2009). These results are related to findings by Cohen (1981) in his meta-analysis study which indicated instructor skills (r = .50) and class structure (r = .47) were significantly related to overall score.

Although student perception has been examined by a large number of researchers during the past fifty years, there is very little published research available on student perception of GEP courses. Qualitative studies have reviewed the composition and purpose of GEP, but only one quantitative study, done at Duke University investigated student satisfaction scores.

Undergraduates are exposed to the GEP which makes it a crucial part of every student's education. This study focuses the lens on a group of students that are just starting their college experience and taking classes that are fulfilling graduation requirements. Analyzing the values recorded on the student questionnaire is an important step in determining student perception of GEP courses. What students identify as key factors (items) in determining quality of instruction will clarify our understanding of the student/teacher relationship. The information obtained by analyzing student perception of instruction information can lead to changes in critical aspects of course presentation and course structure by determining what items are highly related to a student’s overall perception score. Although a significant number of factors associated with individual students, teachers, and courses are not considered due to confidentiality issues, a basic understanding of the relationship between SET items and student response will be explored.

Methodology

All student evaluation information from all GEP courses at a large public metropolitan university in the southeast U.S. for fall 2002 through spring 2009 semesters was used for data analysis. The data set contained all sixteen individual item responses from the student evaulation form completed by students at the end of a course and additional variables generated from course characteristics. The evaluation form used a 1 to 5 Likert scale (e.g., Excellent, Very Good, Good, Fair, Poor) for the sixteen items listed on the form. In addition to the item responses available from each form, information on class size, course year, foundational area, and teaching mode (e.g., face-to-face, web assisted, and web based) was considered in the analysis. Individual courses were grouped by foundational area; this imposed data structure was initiated in order to be as confident as possible that no individual course, instructor or student was identified.

Data screening methods were used to assess data problems that may be due to processing errors or internal consistency problems. Data quality problems can occur when students do not take their task of form completion seriously. Forms consisting of extremely high or low overall evaluation score but having individual items in opposition to the overall score were identified; these forms were removed from the analysis.

Traditional statistical methods and data mining methodology were employed in analyzing the data. Variable association measures determined the relationships between individual items on the form and the overall instructor rating. Decision tree analysis was used to find a model for determining an “Excellent” overall rating or a “Poor” overall rating. In order to provide an honest estimate of the model results, the data set was divided into a subset (approximately 70%) to produce association rules and a separate subset (approximately 30%) to validate the model results (Wang, 2007). There was no sampling procedure applied as part of the process; data mining is specifically suited for searching in large data sets with a large number of variables and missing values (Breiman et al., 1984). No missing data imputation was necessary as data mining techniques process the missing category as a separate group. Missing values provide information for data mining by considering not only the amount of missing data points but the pattern of missing data values. The advantage of data mining techniques is that all data can be used including the missing category, where traditional statistics procedures typically eliminate missing data (Hand, Mannila, & Smyth, 2001). Decision trees are a flexible modeling procedure that results in easily interpretable if-then statements.

Delimitations

This research was conducted with the following delimitation:

1. The study focused on undergraduate GEP courses from Fall 2002 through Spring 2009 from the participating university.

Limitations

Student information was collected prior to the study and not specifically for the purposes of this study. Procedural conditions were not controlled and direct contact with participants was not possible. The data set available was not a random sample of students or courses. Information was collected from students willing to complete the questionnaire; this group of students may represent a biased sample. Under the current data collection procedures, it was not possible to obtain random samples or complete information from all students enrolled in the course.

There may be important components of the variation in evaluation scores that were not being considered as part of this study. Individual student information was not obtained at the time of data collection because the forms are anonymous. Faculty information could be obtained, but for reasons of confidentiality, instructor information was not used in this study.

Information on how students interpret the items on the student evaluation form and what specific instructor actions were being applied to that item were not available at the present time. Thus, an item such as organization of the course is based on the individual student perspective of that particular item.

Results could be generalized to other student populations at similar universities or to future semesters at the same university. Results should be most closely related to future semesters at the same institution, as the students, instructors, and courses will continue to be very closely related to the information used in the analysis. The novel methodology developed here however, should transfer to other institutions with similar student evaluation information.

Assumptions

This study assumes that students were answering truthfully and accurately to the individual items presented on the form. Results could be invalidated if students did not fill out the form seriously, felt pressured to slant the scores, or there was a violation of procedure protocol. No provisions are currently in place to identify any of these possible threats.

It is also assumed that procedural guidelines were followed or that any deviations would not be related to changes in evaluation scores. Aggregation of individual courses allows the focus of analysis to be at the foundational area level.

Summary

Student evaluation information is collected at a majority of colleges and universities in the United States; this information is then used for a variety of purposes. These purposes include: tools for instructional improvement, evidence for promotion and tenure decisions, student course selection, program effectiveness, departmental instructional merit, and possibly teaching awards (Abrami, Theall, & Mets, 2001; Kulik, 2001). Because of the emphasis placed on student scores, it is important to understand, as much as possible, how students arrive at their overall perception score.

The purpose of this study was to understand, for GEP courses, what factors are predictive of the overall instructor rating value given by students. Individual items on the student evaluation form and class size, percent responding, foundational area, and GEP theme were the variables used for modeling the overall instructor score. The study examined what instructor/course qualities were related to a high overall evaluation or a low overall evaluation value.

Data mining techniques are perfectly suited for secondary data analysis. Data mining methodology, specifically decision trees were incorporated to understand the relationship of overall instructor scores with other items and course factors. Decision trees are very efficient modeling tools in situations with large data sets where there are multiple variables with missing values. Missing values do not need to be excluded or imputed in order for modeling procedures to find important variables that will differentiate values of the overall evaluation score.

Prediction models are constructed using a top-down approach; data responses are subdivided and subsequent divisions are based on the previous subgroups. Model results are in the form of easily related if-then-rules for interpreting values of the overall evaluation score. The rules can be assessed using accuracy information in the form of odds ratios and misclassification rates.

CHAPTER 2
LITERATURE REVIEW

The first student evaluation form implemented was most likely at Purdue University in 1927. Almost immediately, Purdue researchers began examining results from these evaluations (Centra, 1993). “No method of evaluating college teaching has been researched more than student evaluations, with well over 2,000 studies referenced in the ERIC system” (p. 495). This quote from John A. Centra (2003) indicates the importance of understanding student evaluation information. The initial review of literature focus for this study was limited to North American higher education institutions. Emphasis was placed on research using some formal student evaluation protocol, published from 2000 -2009, but included previous research by principal investigators commonly cited in recent studies. Abrami, d’Apollonia, Centra, Feldman, Marsh, Roche, and Seldin are commonly cited authors on the topic of student evaluation. Although most studies from the past decade have primarily focused on the association of grades with student evaluation scores, this relationship was commonly debated in the educational community.

Chapter 2 includes seven sections that encompass areas of research pertinent to this study: student evaluation of instruction, structure of evaluation, factors related to evaluation scores, reliability and validity, GEP, data mining, and summary. Section one, student evaluation of instruction, reviews results regarding the specific structure or composition of the student evaluation form along with procedural issues, questionnaire items, and the stated purpose of the evaluation. Section two examines how the structure or content of the information influences results. Section three, factors affecting evaluation scores, is sub-divided into three sections comprising the main factor categories related to student evaluation research: student factors, instructor factors, and course factors. The following section, reliability and validity, highlights research results regarding the relationship of student scores with other measures of effective teaching. The GEP and data mining sections introduce how these components of the study are represented in published research. Finally, a summary of published literature framed for this study is presented.

Student Evaluation of Instruction

In university settings, faculty use assessment methods to evaluate student performance in their courses, so it is logical that students should have some measure of reciprocity. Yet there is disagreement among faculty and administrators on how this information should be used (Abrami, 2001). In face-to-face classes, students spend hours observing the instructor during class meeting times and may have contact with them outside of class or during office hours. Even with this typical student/teacher interaction, there are still aspects of instruction, such as planning, or scholarship that are not perceived by students (Saroyan & Amundsen, 2001). Faculty members may have additional responsibilities such as research and service, which also factor into their overall job performance (Campbell & Bozeman, 2008). How evaluation information is gathered and what weight (i.e., importance) is attached to various aspects is determined by individuals or committees in charge of evaluation (Abrami, 2001).

Evaluation of instruction is a very complex task, not unlike the process of teaching itself. Shao, Anderson, and Newsome (2007) assert that teaching effectiveness can be measured by student evaluations, written comments, peer classroom visits, portfolios, teaching awards, student learning outcome, and scholarship activities. Kulik (2001) commented that many aspects of quality teaching are not observable. Some critics believe that teaching cannot be measured, because there is no consensus for what is “good” teaching. Ornstein (1995) stated that, “Research on teaching was often atomized into tiny behaviors, methods, and/or processes while ignoring the whole picture, that is, larger patterns or relationships of teaching and learning" (p. 3). How is teaching measured? According to Apodaca and Grad (2005), comprehensive evaluation of instruction should include the planning, performance, and evaluation stages.

Analysis of student evaluation information became a major focus in the 1970's when H.W. Marsh began constructing an evaluation form, Student Evaluation of Educational Quality (SEEQ) that identified nine specific components of effective teaching. These components are termed: learning/value, enthusiasm, organization, group interaction, individual rapport, breadth of coverage, examinations/grading, assignments, and workload/difficulty. Marsh and Roche (1997) assert that teaching is multidimensional and that any evaluation instrument needs to adequately address this issue. Other researchers have found similar but not identical components representing the multidimensionality of instruction (Feldman, 1976; Kim, Damewood, & Hodge, 2000). However, not all researchers agree with this assessment but instead contend that student ratings predominately assess a global or general perception (Abrami & d'Apollonia, 1991; Greenwald & Gillmore, 1997; McKeachie, 1997). Cashin and Downey (1992) studied 105 institutions and found that global items were appropriate summary measures of evaluation forms. In their study three control variables were used: class size, difficulty, and student motivation. D'Apollonia and Abrami (1997) prefer a unidimensional approach when information is used in decision making. McKeachie (1997) agrees that the unidimensional approach is practical but goes on to stipulate that any numeric score should only be representative of "crude" assessment categories- "promote" or "don't promote."

There is no denying the importance of student's role in evaluating instruction in higher education. D'Apollonia and Abrami (1997), state that 98% of higher education institutions in the United States use some form of student evaluation and that an increasing percentage of international institutions are doing so as well (Moore & Kuol, 2005). Student information is used for student course selection, promotion and tenure, teaching awards, program effectiveness, school accreditation, and faculty improvement of teaching (Abrami, 2001; Kulik, 2001). In the academic world, student evaluation has been the center of intense debate (Abrami, 2001). Although evaluation of performance is common in most employment environments as a catalyst for improvement, questions regarding how these ratings are related to improvement of instruction have not been adequately explored (Wang et al., 2009).

One central disagreement regarding student evaluation information is the question regarding what specifically is being measured. Theall and Franklin (2001) comment on the student, who may be considered less qualified, determining the quality of instruction. The authors believe that the typical student evaluation scores may show student satisfaction or dissatisfaction but that dissatisfaction may not be equivalent to the instructor not doing an adequate job. Beyers (2008) believes student scores are measuring an "emotional experience" that may not relate to classroom reality. Instructors sometimes feel threatened by the student as evaluator approach (Campbell, 2005). Ory and Ryan (2001) believe that students are in the best position to evaluate an instructor. Preferably, student evaluation data should be collected as one small part of the teacher/course picture which is used to understand the instructional component (Campbell, 2005).

Structure of Evaluation

Evaluation information is obtained using specific methods and procedures, which affect the data quality and ultimately the intended use of the information (Hand, Mannila, & Smyth, 2001). Three important components of gathering evaluation information are: procedure, structure, and purpose. Procedures are the exact conditions associated with the collection period.

Evaluations whether based on interview, focus groups, or questionnaires have specific items or questions that are addressed. Understanding how these design components interact with the evaluation process is paramount, especially for faculty being evaluated and evaluators using the information. Structure refers to the methods used to elicit student information.

Information from students can come from interviews, focus groups, personal contact, web sources, or questionnaires. Most universities use individual student questionnaires as the primary method for collecting information (Seldin, 1999). Using a pre-determined questionnaire allows each teacher/course to be evaluated based on identical components. Not all evaluation forms are created equal. Marsh and Roche (1997) voice concerns regarding the usefulness of forms comprised of poorly worded or inappropriate items. When different courses, questionnaires, procedures, and institutions are added into the picture, it becomes very difficult to make authentic comparisons using previously published empirical studies.

The structure of the data collected should be directly connected to the purpose (Algozzine et al., 2004). Questions or items on a questionnaire could be free-response, numerical, or categorical making the appropriate summary and analysis different for each type of response. Having a clear focus or intended purpose for the resulting information prior to collecting the information would be preferable. But due to the nature of complex systems, data by the very nature of being available tends to get used by multiple groups for a wide variety of purposes (Hand, Mannila, & Smyth, 2001).

Is it fair for a seminar class or a web-assisted class to have the same items evaluated as a face-to-face class? Abrami, Theall, and Mets (2001) report that the instructional mode or delivery is changing from traditional didactic forms of instruction to more learner-centered approaches. Forms that ask students to evaluate characteristics associated with face-to-face classroom settings may not be appropriate for a course delivered via the web. Theall and Franklin (2001) reiterate the importance of a clearly defined policy and process for student evaluation information.

Committees designing questionnaires cannot anticipate the multiple uses of collected information for unintended purposes. These undefined uses of the information open the door for misuse, according to Seldin (1984). In order to combat misunderstanding of student results, Theall and Franklin (2001) recommend establishing guidelines that accommodate the concerns of all stakeholders.

Evaluation Procedure

Collection procedures need to be included as part of the complete picture, prior to dissemination of information (Hand, Mannila, & Smyth, 2001). Where and how is information collected? D'Apollonia and Abrami (1997) found that administrative conditions such as student anonymity, proctoring, and the stated purpose of the evaluation, need to be standardized. Wachtel (1998) believes that anonymity, instructor presence, stated purpose, and timing of the evaluation are all potential influencing factors. Feldman (1979) found that student ratings are higher when the instructor is present. D'Apollonia and Abrami (1997) conclude that the relationship between student learning and student ratings is significantly higher when the evaluation is carried out after rather than before the final examination. These are the only experimental studies regarding procedural issues and the relationship with student evaluation scores.

There are conflicting empirical studies regarding the influence of procedural deviations. Procedural violations may be done purposefully or inadvertently; examples of tactics used are: letting students out early, administering the evaluation after a fun activity, remaining in the room, providing bonus points, or picking a day when certain students are absent (Pounder, 2007). Anecdotal evidence can be obtained by asking almost any faculty member about problems with student evaluations (Beyers, 2008). However, D'Apollonia and Abrami (1997) found no significant evidence that procedural issues influenced the validity of the results.

As web-based contact with students increases it is probable that student evaluations will move to the online environment and this move is likely to change the completion rate and the type of students motivated to complete the evaluation. In fact, recent studies have hypothesized a polarizing effect when instructor evaluations are completed on line (Benton, 2008). Yet a study done by John Goyder (2009) at University of Washington found only a mere suggestion of this effect.

Items or Questions on the Evaluation

Identification of the components to be measured is the first step in creating items for a questionnaire. What is the questionnaire attempting to measure? Ory and Ryan (2001) do not believe students should be asked to evaluate course content. Theall and Franklin (2001) state that, "beginning students do not have sufficient depth of understanding to accurately rate the instructor's knowledge of the subject" (p. 49). Herbert Marsh who has published extensively in the area of student evaluations; he says that questionnaires need to address the multidimensionality of teaching (Marsh & Roche, 1997). Marsh developed a form for student perception responses based on his nine dimensions of effective teaching. These dimensions are: learning/value, enthusiasm, organization, group interaction, and individual rapport, breath of coverage, exams /grading, assignments, and workload/difficulty. Abrami and d'Apollonia (1991) use rapport, interaction, feedback, evaluation, and difficulty as the dimensions of interest. Chickering and Gamson (1987) have seven principles which emphasize: contacts between students and faculty, developing reciprocity and cooperation between students, using active learning techniques, prompt feedback, time on task, high expectations, and respecting diverse talents and ways of learning. Seldin (1984) used previous student evaluation studies to identify the following teaching components: “being well prepared for class, demonstrating comprehensive subject knowledge, motivating students, being fair and reasonable in managing the details of learning, and being sincerely interested in the subject manner and in teaching itself” (p. 133).

There is no standardized set of items that evaluate effective teaching and empirical research studies use different sets of items in their analysis. Kolitch and Dean (1999) examined items typically given on student evaluation instruments and determined the forms were not broad enough to measure all aspects of instruction. These differences in items and forms create an environment where generalizations are misleading. Differential weighting of items can exacerbate the issue of comparability.

Multidimensionality of teaching, from a student perspective, is not an agreed upon concept. Apodaca and Grad (2005) found that student ratings could be interpreted as multidimensional as much as unidimensional. Marsh (1984) believes that if only one value is used, it should be a weighted value. One-dimensional or global measures of teaching are defended as the preferred measure used in decision making (Abrami & d’Apollonia, 1991). Due to the individuality of student evaluation forms, conclusions regarding global measures are varied.

Studies on whether items should be open-ended or closed and if closed, what scale should be used, dictate that item construction should be related to purpose (Dillman, 2000). Open or free response questions are recommended when detailed or formative evaluation information is collected (Dillman, 2000). Closed questions have an assigned scale for providing answers. Dillman states that closed-ended items are, “most useful when one has a well-defined concept for which an evaluative response is wanted, unencumbered by thoughts of alternative or competing ideas” (p. 43-44).

Purpose of student evaluations

Student evaluations can be used for formative assessment or for summative assessment. Timing, content, and structure needs to be compatible with purpose. Formative evaluation is extremely useful for faculty just starting their careers, when they are still developing their personal style of teaching (Kulik, 2001). For formative evaluation, a multidimensional use of the information is more appropriate, given the goal of improved instruction. In this context, different aspects of teaching (e.g., organization, delivery, assessment) can be critiqued separately. The evaluation, whether summative or formative, may be used by the faculty member to make changes in their teaching process or it may be viewed by students, peers, or administrators.

Questionnaires presented to students at the close of a course are typically utilized for summative assessment. Summative assessment is useful for evaluation purposes; it is more practical to use one overall score. Global rating scores or overall rating values have been widely studied and the generalizability of one score is less problematic (Abrami & d’Apollonia, 1991). Franklin and Theall (1989) and Abrami (2001) comment on the difficulty administrators would have properly weighting multiple scores in arriving at a decision regarding quality of instruction.

When students are not informed of the evaluation purpose they may not be careful completing the form (Campbell & Bozeman, 2008). Aleamoni and Hexner (1980) found that students informed that the purpose of the evaluations was for administrative purposes rated instructors higher than when the instructor would be the only person viewing results.

Factors Related to Evaluation Scores

Understanding what factors might be related to evaluation scores is important for faculty being evaluated and for people using the evaluation information. Too often the scores are misunderstood or misused in evaluating teaching (Kulik, 2001). Individual faculty scores are not as controversial as comparing scores across teachers, class level (undergraduate or graduate), class type (GEP, required, or elective), disciplines (Arts, English, Engineering, Natural Science, Social Science, etc…), class size, or other differences. Unfortunately, empirical studies give conflicting results in regards to significant influences because the studies are isolated in a single university using different sets of characteristics. Factors that might show differences in evaluation scores given by students fall into three categories: student, teacher, and course characteristics (Pounder, 2007).

Biasing factors are defined as influences external to the true teaching environment. Biasing factors are traits that are related to student evaluation scores even though the traits do not have any theoretical relationship with effective teaching or student learning. Fixed instructor traits such as age, gender, and race that cannot be changed are examples of student held biasing factors (Sprinkle, 2008). Characteristics that influence student ratings as a result of those characteristics being directly related to teaching effectiveness should not be considered a biasing factor (Marsh, 1983). An example of a characteristic that could be influential yet not a biasing factor would be when student learning is related to class size.

Student Factors

Student factors are characteristics of the individual student such as gender, race, ethnicity, age, major, and interest. Empirical studies have not consistently indicated any set of student characteristics to be significantly related to evaluation scores. This may partially be due to the lack of data collection on student attributes. A systematic review of sixty-eight studies done by Pounder (2007), found a few studies (Bachen, McLoughlin, & Garcia, 1999; Feldman, 1993; Walembwa, Wu, & Ojode, 2004) reporting a gender effect, yet the relationship was not consistent across studies. Other studies reporting gender differences are not cited due to small sample sizes. Marsh (1983) found four characteristics that were associated with higher ratings; one of those characteristics was student interest.

Other studies have looked at a variety of student factors such as effort, locus of control, ethnicity, and learning style. Heckert, Latier, Ringwald-Burton, and Drazen (2006) found student effort was positively related to all dimensions of course evaluation. Feldman (1989) also found student motivation to be related to student achievement. Grimes, Millea, and Woodruff (2004) used locus of control as a measure of internally or externally oriented student beliefs. The authors found that students with more internally oriented locus of control score were more likely to assign above average evaluation marks for instructor performance. Qualitative studies have shown that certain ethnic groups may have difficulty rating professors because of cultural factors (Ory & Ryan, 2001). Hativa and Birenbaum (2000) found that the student’s specific learning style was directly related to their definition of good instruction. Crumbley, Henry, and Kratchman (2001) discovered from interviews that students use evaluations to punish instructors who ask embarrassing questions, give quizzes, or are tough graders.

Instructor Factors

Instructor factors are similar to student characteristics already described. In addition to demographic information, teacher characteristics include personality identifiers and classroom management characteristics. Instructor rank, organization skills, sense of humor, student rapport, content delivery, and test construction are all examples of additional teacher characteristics. Feldman (1993) summarized results from ten studies and found very little support for gender differences but when there was a significant difference, typically, the female instructors scored higher. Research on instructor rank, age, gender, and experience has produced mixed results (Bachen, McLoughlin, & Garcia, 1999; Feldman, 1993). Gravestock and Gregor-Greenleaf (2008) did not find that rank or experience had a measureable impact on ratings. Radmacher and Martin (2001) found teacher extraversion was the only significant predictor when modeling student evaluation scores. Other variables in their model included grades, student age, and class size; this information did not contribute unique information to the model. Clayson and Sheffet (2006) and Feldman (1986) identified a significant relationship between instructor personality and student scores although these personality traits may not have any true educational value. Few large sample empirical studies use student or instructor characteristics due to confidentiality issues.

Course Factors

Courses have specific characteristics that may have an influence on student evaluation scores. Type of course includes course level (undergraduate or graduate), course type (GEP, required, or elective), course mode (face-to-face, web-enhanced, or fully on-line), and course discipline. There are also course components such as work load, laboratory requirements, or imposed grade structures that can be related to evaluation scores. A major focus of student evaluation research in the last twenty years has been the relationship of grades and student evaluation scores. Additionally, there may be characteristics of the student or faculty population that are unique to a specific university, these characteristics may interact with the course characteristics differently at one university compared to another university.

Greenwald and Gillmore (1997) examined grades and workload in relation to instructor ratings. Centra (2003) found courses rated by students as the “right” difficulty level received higher evaluations than either too difficult or too elementary. Eiszler (2002) found the relationship between student ratings and expected grades was significant even after controlling for possible alternative effects. Pounder (2007) in his review of literature found numerous studies verifying the relationship between expected grades and evaluation scores. Although the novice instructor may believe that giving higher grades will translate into higher scores; that is not always the case. McKeachie (1997) indicates that the effect of easy grading may vary by institution. In a Canadian study by Gravestock and Gregor-Greenleaf (2008) the authors found a positive correlation between grades and ratings. Griffin (2004) using regression analysis concluded that instructor leniency, as perceived by students, was positively associated with student ratings.

Chang (2000) found, using Taiwan Teachers College data that, student enthusiasm, participation, expected grade, grading, and course difficulty were significant variables for a regression model predicting an overall score. Similarly, Remedios and Lieberman (2008) found that student perceived quality of teaching was related to how much students enjoyed or felt stimulated by the course content.

Previous research results generally have found that students rate elective courses higher than required courses (Brandenburg, Slinde, & Batista, 1977; Costin, Greenough, & Menges, 1971; Dooris, 1997; Feldman, 1978). More recently, Aleamoni (1999) agreed with the previous results where a negative relationship between scores and required classes was found. The previously referenced studies represent the most recent published results relating elective or required courses to student evaluation information.

The academic discipline of the course may also have an influence on the student evaluation scores. Mathematics and natural science courses consistently exhibit lower scores whereas literature and history have higher scores (Cashin & Clegg, 1987; Feldman, 1978). This relationship may be even more evident when mathematics courses are required GEP courses. Cashin and Clegg (1987) and Feldman (1978) represent the most recent credible published results regarding course area and student evaluation scores.

McKeachie (1997) contends, “There is ample evidence that most teachers teach better in small classes” (p. 1220). Marsh, Overall, and Kesler (1979b) found the class size effect depended on the specific components of the evaluation being administered. Overall scores did not change relative to class size, whereas student/teacher interaction item was adversely affected in larger classes. Their study had class sizes from 5 to 409 students, from six different departments, representing undergraduate classes at the University of California, Los Angeles (UCLA). Centra (2000) determined that classes with less than fifteen students get higher evaluation scores. Koh and Tan (1997) and Toby (1993) concluded that class size was related to evaluation scores and that as class size increased the scores decreased. Centra (2003) interpreted results of class size studies, that if small classes produce higher scores on student evaluations because students are learning more, then class size is not truly a biasing factor.

Abrami, Theall, and Mets (2001) reported on changes in instructional mode and how those changes should be considered when using student evaluation information gathered from multiple types of instructional environments. The authors concluded some evaluation items may not be appropriate for a fully online course. Additionally, a student's ability to work with technology or apprehension of this form of interaction may influence their perception of the course (Abrami, Theall, & Mets, 2001). Summers, Waigandt, and Whittaker (2005) found that students enrolled in online classes were significantly less satisfied with the course compared to the face-to-face version of the same course.

Reliability and Validity

Student ratings are used at a majority of American colleges and many researchers view the results as reliable and valid measures of teaching (Kulik, 2001; Marsh & Roche, 1997). The major problem with any source of information on instruction is no one has an extensively accepted criterion for effective teaching (Kulik, 2001). Without a common goal, it is difficult to find agreement among experts on what measures should be used to evaluate effective teaching. As with most controversial topics, there are conflicting and confusing results in published studies.

Reliability is consistency. Would the scores given by these students change if given at a different time or under different conditions? Marsh (2007) and Carle (2009) found student evaluation scores to be stable over time. Many studies have found student ratings obtained at the close of the course to be positively correlated with alumni ratings, which are collected after students graduate (Feldman, 1989; McKeachie, 1987; Overall & Marsh, 1980).

Validity is determined not from the form itself but from how the information is used or how the scores correlate to other sources. These sources could be such items as: perceived student learning, student comments, standardized tests, peer reviews, alumni ratings, and expert reviews (Kulik, 2001). McKeachie (1997) does not believe that the specific items used or potential biasing factors are a major threat to validity. For student evaluation scores to be valid, these scores need to show a positive relationship with other criteria believed to measure effective teaching (Kulik, 2001). Cohen (1982) states, “most researchers in this area have agreed that student learning is the most important criterion of teaching effectiveness” (p.78). Roche and Marsh (2000) found the measure of perceived learning and overall score have a correlation of 0.53 (286 courses) which is interpreted, by the authors, as a large effect size. Also statistically significant was the correlation between engaging assignments and an overall score. Marsh’s results, given above, use an overall score computed from the SEEQ form, developed by Marsh in 1982. Cohen (1981) using meta-analysis to examine forty-one multi-section courses, found a statistically significant correlation of .43 for student achievement and overall score.

Ory, Braskamp, and Pieper (1980) found that student comments from interviews, written responses, and ratings scores gave similar pictures of teaching effectiveness. Meta-analytic reviews by Cohen (1981, 1982) and a review of forty studies by Feldman (1989) clearly show that examination scores are related to rating averages (Kulik, 2001).

Marsh (1984) found that trained observer scores were correlated with student scores but peer scores were not related to student ratings. It may be that peers were judging the classroom aspect for a limited period only, which may be very different from the student experience because students are interacting with the faculty member over the entire course both in the classroom and outside of the classroom (Kulik, 2001). Noting that trained expert scores are similar to student scores while peer scores are not, may indicate peers were making judgments from a teacher perspective and not a student perspective (Kulik, 2001). Alumni ratings usually have a very low response rate and thus are not a source of credible information (Scriven, 1983).

Critics of student evaluations believe that scores students give are related to the grade they received or the grade they expect to get and thus are not really reflective of the quality of teaching. One method of checking this belief is to look at the relationship of grades and evaluation scores. Marsh and Roche (1997) compiling 9,194 course section results from various institutions, in a variety of course disciplines, estimate the average correlation between the overall student score and anticipated grade to be approximately 0.20. Although the authors agree this correlation value is statistically significant and could be a biasing factor, they interpret the value as predictable because grades should reflect learning.

Validity of student ratings has been studied using multisection studies, multitrait-multimethod studies, external influences, laboratory designs, and dimensionality framework (Greenwald, 1997; Ory & Ryan, 2001). Multisection analysis uses the same course and the same exams but different instructors. Cohen (1981) used forty-one previous studies and concluded that some studies had high positive correlations and other studies had negative correlations between instructor scores and test scores. He concluded that different items, full-time instructors, and the timing of the evaluation may have contributed to the conflicting results.

Abrami, d'Apollonia, and Cohen (1990) comment, "In one view, student ratings are valid if they accurately reflect students' opinions about the quality of instruction, regardless of whether ratings reflect what students learn" (p. 219). They conclude that at a minimum student scores represent student satisfaction.

Previous GEP course analysis at Duke University has shown that student differentiations, item to item differences, are a valid measure of understanding teaching and learning components (Thompson Jr. & Serra, 2005). Findings in the Duke study show relational agreement between faculty and student perception of GEP course characteristics. Marsh, Overall, and Kesler (1979a) also found agreement between instructor and student perception (*r* = 0.49). An example would be when instructors rate a course low on the “gaining factual knowledge” scale, the student response is also low.

General Education Program Evaluation

General education was first conceptualized at Yale College in 1828 and slowly developed into a set of courses representing the basics or essential components of education (Awbrey, 2005). By 2003, approximately ninety-five percent of four-year colleges and universities had some type of GEP designated for undergraduates (Aloi, Gardner, & Lusher, 2003). Institutions with a GEP or core curriculum have a designated set of classes that need to be taken prior to graduation. Examples of courses in the GEP include classes like college algebra, English composition, and world history. The 1980’s was a period of intense GEP reform due to internal and external catalysts. Institutions of higher education were experiencing declining enrollments, sagging reputations, increase in faculty input regarding aspects of student education, and having to meet accreditation standards (Awbrey, 2005). Johnson (2000) reported the most common change in general education requirements during the 1990’s was structural. Thematic organization and interdisciplinary sequences were the primary strategies for changing the GEP curriculum specifically to bring more cohesion to the program. Glynn, Aultman, and Owens (2005) emphasize general education should incorporate cohesion, integration, and interdisciplinary connections throughout the curriculum. The current focus on GEP programs and programs in general, is on student learning, this accountability is influencing GEP program assessment (Joe, Harmes, & Barry, 2008).

Although student perception has been examined by a large number of researchers during the past fifty years, there is very little published research available on student perception of GEP courses. One of the few studies was done by Duke University, Trinity College, Office of Assessment (TCOA). This office is in charge of all student evaluation information for the university. Thompson, Jr. and Serra (2005) published results of a study addressing student learning objectives of their GEP courses and the relationship to student satisfaction scores. The objective of the study was to align the student course evaluation process with learning objectives for the course. Results from 1,100 course sections showed student and faculty had a general agreement when rating various aspects of the course. Courses with different learning objectives were rated differently and courses with the highest intellectual stimulation were rated as high quality courses (Thompson Jr. & Serra, 2005).

Data Mining

Data mining has recently been added as an alternate analysis technique to the study of student evaluations and educational data but few studies have been published. Wang et al. (2009) published results generated from a data set containing all courses from the same public institution for academic years 1996-2001. The researchers used data mining techniques to develop a model for overall evaluation rating using 15 individual items, course level (lower undergraduate, upper undergraduate, or graduate), and academic year (1996-97, 1997-98, or 1998-99). The decision tree results focused on three qualities of effective teaching: effective communication skills, facilitative teaching, and organization and assessment (Wang et al., 2009). No other factors (course level, college, and year) were found to be important for modeling.

Thomas and Galambos (2004) used decision tree modeling to investigate factors related to overall student satisfaction using a seventy-nine question satisfaction survey given to 1,698 students at a public university. The questionnaire covered topics related to the entire college experience with no items specifically addressing individual courses. The results indicated that faculty preparedness was the only classroom factor related to general student satisfaction regarding their college experience. Student demographics were not significant predictors of general student satisfaction. Faculty preparedness is similar to the organization quality found by Wang et al. (2009).

Summary

Student evaluation information is readily available and multiple groups (students, instructors, administrators) use this type of information to make a wide variety of decisions. Although student satisfaction scores do not measure student learning, research studies generally find a positive correlation between instructors who receive high ratings and some measure related to effective teaching (Kulik, 2001). Centra (2003) concludes that student evaluation information can be used to improve instruction.

Unfortunately, a universally accepted definition of effective teaching or instruction is not known (Trout, 2000). The construct of effective teaching, changes based on the specific circumstances surrounding the instructional process. However, there is a general agreement among experts in the field of student evaluation, that at a minimum, these scores represent student satisfaction (Abrami, d'Apollonia, & Cohen, 1990). Additionally, there is substantial research linking student satisfaction ratings to effective teaching (Theall & Franklin, 2001).

Empirical studies use class, student, and instructor characteristics as a gateway to understanding what factors are identified as important, from the student perspective, in the variability of numerical scores collected from student evaluation forms. In regards to published research, there is no conformity across studies, regarding a set of student, teacher, or course factors that consistently relate to evaluation scores (Ory & Ryan, 2001). Centra (2003) concluded that student evaluation scores are only minimally affected by instructor, student, or course characteristics.

Required courses, larger classes, mathematics courses, and natural science courses typically show lower average instructor evaluation scores (Feldman, 1978; Theall & Franklin, 2001). Many other student, course, or instructor factors have shown conflicting results in studies isolated in specific disciplines of individual institutions (for a review of sixty-eight published studies see Pounder, 2007). Conflicting results are commonplace due to the complex structure of student evaluation information. Different questions, form construction, and procedural instructions also add to the complexity of information obtained.

It is also interesting to note that some of the highly publicized studies showing effects of grade inflation, personality characteristics, and other physical characteristics of the instructor, are highly flawed (Kulik, 2001). These studies had very low sample sizes or used designs, such as viewing an instructor video clip for 30 seconds without any sound, that were not representative of a true instructional environment.

Data collection issues limit most studies to one institution. The unique characteristics of that setting may not generalize to other institutions especially considering the different methods (e.g., questionnaires, focus groups, or interviews) used to collect student information. The purpose, and thus the details surrounding what type of information will be collected are institution specific. These different characteristics make each study unique; consequently the findings for that institution may not carry over to other institutions. Student evaluation information is misused and will continue to be misinterpreted unless institutions become aware of the quality and value of the information (Theall & Franklin, 2001). McKeachie (1987) stated, “In many institutions the problem of evaluation is not a lack of data but rather the ineffective use of the data available" (p. 348). It is imperative that each institution develop a rationale for using the evaluation information (Ory & Ryan, 2001).

CHAPTER 3
METHODOLOGY

The purpose of this study was to determine, for GEP courses, what individual items on the student evaluation form and course variables were predictive of the overall instructor rating value. Using data mining techniques, instructor/course qualities that related to a high overall evaluation or a low overall evaluation score were identified. Additional information on course year, instructional mode, GEP foundational area, and class size was examined in relation to overall scores given by students. Data mining techniques provided relatively new tool for understanding student evaluation scores.

The research questions and information used in this study are presented in Chapter 3. Next, a section describing instrumentation is followed by a description of the variables of interest. The chapter continues with data screening, data construction, and ethical issues. The data analysis section provides information on analysis techniques employed to answer the posed research questions.

Research Questions

1. What items on the student evaluation form are related to the “Overall rating” (item 16) score provided by students?
2. Is instructional mode, class size, GEP foundational area, or incorporating a GEP theme related to the “Overall rating” (item 16) score provided by students?
3. What rules can be identified to understand the determination of an “Excellent” overall rating on the student evaluation form?
4. What rules can be identified to understand the determination of a “Poor” overall rating on the student evaluation form?

Data Collection

Data utilized in this study were previously collected by a large public metropolitan institution of higher learning for other purposes. All observations from the present student questionnaire for all GEP courses in consecutive academic years 2002-2003 ending in spring 2009 were used to constitute the data set of interest. Student scores for 43 different courses taught in 8169 sections were provided by Office of Academic Services personnel in text format.

The student questionnaire was a 16 item, Likert-scale, response form that students complete in the final two weeks of the course. Of the sixteen items on the questionnaire, eight were constructed by an external advisory board and the remaining eight items were developed by a university-wide committee (Dziuban, personal conversation, September 2, 2009). Additional prompts on the form were free response items which were not coded in the database and therefore were not available for analysis. Web assisted courses and online courses had an electronic version of the same questionnaire available via a secure portal.

Standard university protocol for student evaluations is as follows: during the final two weeks of each term, forms with procedural instructions attached are distributed to the instructor; these instructions specify that a proctor is to distribute the forms to students during the first 15-minutes of the class period. Forms were assumed to be completely anonymous and immediately collected for computer scanning. No provisions were in place to oversee the actual implementation of procedures for questionnaire completion. Potential procedural variations may have some effect on the resulting scores, but the variation cannot be controlled or identified.

During this period (2002 -2009) on-line classes had the option of conducting the evaluation on-line or via paper. Instructors using the on-line option notified students of the web address for the evaluations. For this study, only a small percentage (12.29%) of GEP courses were conducted using alternative modes (e.g., video feed, reduced seat time, or web-based), thus the number of evaluations submitted on-line is probably limited. How the evaluation was completed, whether on paper or on-line, was not added to the digitally stored data sets.

Only GEP courses available for fall 2002 through spring 2009 were used in this study. The GEP courses were grouped into five areas; the foundational areas of interest in this study were: a) communication, b) cultural and historical, c) mathematics, d) social, and e) science (institution undergraduate course catalog, 2009). Within each foundational area two to four courses must be completed regardless of a student's major. GEP courses are a group of classes taken by undergraduates in order to fulfill graduation requirements. The courses are institution specific and generally cover areas of importance, as designated by the institution (Glynn, Aultman, & Owens, 2005). The philosophy of the GEP is for students to have a broad understanding of diverse disciplines (Awbrey, 2005). The present course catalog gives the philosophy of GEP as, "to introduce students to a broad range of human knowledge and intellectual pursuits, to equip them with the analytic and expressive skills required to engage in those pursuits, to develop their ability to think clearly, and to prepare them for life-long learning" (institution undergraduate course catalog 2009, P. 48).

GEP courses, as a group, are similar in regard to the type of student taking the course; these students are undergraduates, typically in their first few years of higher education. It was hypothesized that these students/courses had similarities that created a more homogeneous population. This cohesiveness created an opportunity to understand the important instructor qualities identified by students taking GEP courses.

Instrumentation

The instrument “Student Perception of Instruction” used in this study contained 16 items measured on a five point Likert scale and is provided in Appendix A. Eight items were common to the state university system and the remaining eight were institution specific. The response options were: Excellent, Very good, Good, Fair, and Poor. The form was designed to provide student feedback to instructors and other stakeholders on aspects of instruction. Previous internal analysis indicated that a one factor solution was optimal (Dziuban, Wang, & Cook, 2004). In other words, the analysis results indicate the form items are measuring one global construct. There are no studies regarding the validity or reliability of the scores produced from this instrument.

Variables and Measures

Data collected for this study was compiled from the student evaluation form and university enrollment records. Variables defined as Area and GEP were generated from SAS code as part of the programs used to merge, condense, and purge information in order to obtain a master data file for analysis. Table 1 gives information pertaining to variable names, labels, values, and scale of measurement.

Table 1

Identification of Variables

Variable	Label	Values	Scale
Q1 to Q15	Individual items	1 = Excellent 2 = Very good 3 = Good 4 = Fair 5 = Poor	Ordinal
Q16	Overall Score	1 = Excellent 2 = Very good 3 = Good 4 = Fair 5 = Poor	Ordinal
Class Size	Number of Students Enrolled	1 to 529	Interval

Area	Foundational Area	C=Communication
		S=Social
		H= Cultural and Historical
		M=Mathematics
		N=Science
			Nominal

GEP	Theme	0=No (2002-2005) 1=Yes (2006-2009)	Binary

Mode	Instructional Mode	F=Face-to-face
		O=Other *
			Nominal

Pctrsp	Percent Responding	1 to 100	Interval

**Other group was composed of email enhanced, video feed, ITV feed, face-to-face with video feed, reduced seat time, tech classroom, two-way television, video stream and web-based.*

Data Screening

Data screening methods were employed to identify data problems that may be due to processing errors or internal consistency problems. Student evaluation information was entered in an electronic data format using a data scanner, so coding problems were minimal. Frequency distributions were examined to verify all coded responses for items 1 to 16 were valid data values (i.e., 1 to 5). If a student did not respond to the overall instructor rating item (Q16) or if a response was missing all items (Q1 to Q15), the entire observation (Q1 to Q16) was removed from further analysis.

Internal consistency problems can occur when students do not take their task of form completion seriously. Any response with an overall instructor evaluation score of “Excellent” or “Poor” had the remaining 15 item responses reviewed in order to identify forms where the responses were clearly not credible. If the response to all individual items was “Poor” or “Excellent” coupled with an “Excellent” or “Poor” (respectively) overall instructor evaluation score, then the form was removed from the analysis.

Class size was checked against the number of students responding to the questionnaire. If the number of students completing the questionnaire was higher than the number of enrolled students, then the entire course was removed from analysis.

Data Set Construction

Student questionnaire responses were processed at the end of each semester. For each academic year, there are three semesters (fall, spring, and summer) with each semester contained in a separate data set. A total of twenty data files, one for each semester (fall 2002 through spring 2009), were supplied from Academic Services containing student response values for each GEP course taught during that semester. A header identifying the college, department, instructor, course, and section, separated each group of student responses. During the merging process, each data file had an additional variable for year/semester added.

In addition to the 16 items on the student form, additional variables were created to identify class size, percent responding, instructional mode, foundational area, and GEP theme. These additional course variables were used to understand what relationship, if any, these characteristics had on students' perceptions. These variables and the values of each were provided in the previous Variables and Measures section. Class size was the number of students officially enrolled in the class section as provided by Institutional Research personnel. Percent responding was calculated as number responding to course evaluations divided by number enrolled. Instructional mode refers to how students interacted with the instructor. Different instructional modes, as designated by the Office of Institutional Research, for GEP classes were: face-to-face, reduced seat time or web-assisted, video stream, and fully on-line. The modes were collapsed into face-to-face or other due to the small number of courses using other modes of instructional delivery. Foundational area was defined to be within the GEP structure (institution course catalog 2008-2009). Five foundational areas used for the academic years 2002-2009 were: communication, cultural/historical, mathematics, social, and science. Each course in the GEP program was designated as part of one foundational area (see Appendix B).

Year/semester information was used to create a binary variable representing GEP theme (yes, no). The first four academic years in the data set (2002-2003, 2003-2004, 2004-2005, 2005-2006) were years prior to the instituted theme of global climate change. The remaining three academic years (2006-2007, 2007-2008, 2008-2009) were with the theme in place. A unifying theme was introduced as a way to integrate information from very diverse courses that, from a student perspective, may not seem related to each other. The current unifying theme instituted in 2006 is “global climatic change” and was selected after extensive student interviews indicated this topic is a major concern to students (Dean of Undergraduate Studies, personal communication, June 4, 2009). An example of thematic integration would be to have students in a core English course compose a paper on a specific prompt around the general theme of global climate change, while at the same time having those same students use data collected on hurricane incidences as a lab activity in their core statistics course.

The student evaluation data was merged with class size and instructional mode information to create one master file, this file was subsequently used for analysis.

Ethical Issues

Students, instructors, courses, colleges, and sections were not identified in any constructed data file used for analysis. All prior data sets were deleted in order to respect privacy issues. Summary information was at the foundational area level only. This imposed data structure was initiated in order to be confident that no individual course, teacher, or student was identified.

Data Analysis

The data set prior to analysis contained sixteen student response scores (Q1-Q16), class size, percent responding, foundational area (communication, cultural/historical, mathematics, social, or science), GEP theme (yes/no), and instructional mode (face-to-face or other). The final data set contained 329,507 student responses. Prior to analysis of the research questions, individual item investigation was conducted. Class size and percent responding were checked by foundational area, mode, and GEP theme variable.

Research question one addresses the relationship of each individual item on the evaluation form with the overall instructor rating item. To answer this research question, Spearman correlation coefficients were computed to measure association because of the ordinal composition of the items contained on the evaluation form. The relationship between the overall rating score (Q16) and the remaining items (Q1 to Q15) were computed in order to understand the association of items on the student evaluation form. Spearman's correlation (r_s) uses ranks to calculate a measure of relationship between two variables; the value range is from $-1 \leq r_s \leq +1$ (McClave & Sincich, 2006). Spearman values closer to either -1 or +1 indicate strong association. The formula for the Spearman's rank correlation value is given below. In the following formula *n* refers to the number of pairs of observations, u_i refers to the rank of the *i*th observation in sample 1, v_i refers to the rank of the *i*th observation in sample 2.

$$r_s = \frac{\sum (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum (u_i - \bar{u})^2 \sum (v_i - \bar{v})^2}} \qquad (1)$$

For research questions two, three, and four, data mining modeling techniques were used to understand the relationship of individual item responses and course variables to the overall score. Research question two asks, "Is instructional mode, class size, GEP foundational area, or incorporating a GEP theme related to the "Overall rating" score provided by students?" Research question three asks, "What rules can be identified to understand the determination of an "Excellent" overall rating on the student evaluation form?" Research question four asks, "What rules can be identified to understand the determination of a "Poor" overall rating on the student evaluation form?"

Data mining is characterized as secondary analysis; this type of analysis is a discovery process (Hand, Mannila, & Smyth, 2001). All available student responses for the group of interest were used in this study for model building. A focus on statistical inferences using sample information to predict population values is not relevant when using population information (Hand, Mannila, & Smyth, 2001). Data mining is specifically suited for searching in large data sets with a large number of variables and missing values (Hand, Mannila, & Smyth, 2001). No missing data imputation was used as data mining techniques process the missing category as a separate group. Missing values provide information for data mining by considering not only the amount of missing data points but the pattern of missing data values. The advantage of data mining techniques is that even missing data as a separate category level can be used, whereas traditional statistics procedures typically eliminate missing data. Decision tree methodology can efficiently handle analysis that would be impossible for logistic regression due to the sparseness of the data values. Another advantage of decision tree results is the user friendly if-then rules, identified through the modeling process. Calculation of odds ratios indicates the usefulness of classification results.

Decision tree analysis was used to find rules for determining an “Excellent” overall rating and a “Poor” overall rating. The target variable was the overall score, item 16 on the form. The rating categories were: Excellent, Very Good, Good, Fair, and Poor. The remaining items (Q1 to Q15) also used the same ordinal scale. Items one to fifteen (Q1 to Q15), class size, percent responding, instructional mode, foundational area, and GEP theme were the independent variables used to find splits to create homogenous groups in relation to the overall evaluation score.

Univariate binary decisions using the independent variables were used to determine subset membership (Hand, Mannila, & Smyth, 2001). The splitting process starts with one binary split based on the most important independent variable; this is termed the parent node which splits into two child nodes. Each subsequent split uses only data observations from the preceding node (i.e., subset). Splits were created that partition data into subsets that are as homogenous as possible as related to the target variable of interest (Breiman et al., 1984).

Final tree size was determined by incorporating two subgroups from the original data set into the model building process. The first data set was used to grow the tree, to a larger than optimum size, while the second data set prunes or cuts back the tree. In order to provide an honest estimate of the modeling results, the data set was divided into a subset (approximately 70%) to produce association rules and a separate subset (approximately 30%) to adjust the model complexity (Wang, 2007). The final model identifies association rules for describing variable relationships with the target variable (overall scores). This data set division was based on a stratified sample of student responses using foundational area, class size, GEP theme, and mode as the stratification criteria. Stratification provides a method that creates representative sub-groups for the analysis process (Breiman et al., 1984).

Model adjustments can be done manually or specified using software options. Change in misclassification rate was used to verify model complexity. The goal is to have a model picking up true differences and not modeling extraneous information (i.e., noise). Another check of model validity was conducted by calculating the rule performance percentage for each foundational group separately. A logical check of model applicability is to assess items appearing in the rules with respect to research on teaching excellence.

Summary

In this study, student evaluation scores from all GEP courses in consecutive academic years 2002-2003 ending spring 2009 were used as the data set of interest. This information consisted of sixteen items measured on an ordinal scale with values from 1 to 5 representing categories of (Excellent, Very good, Good, Fair and Poor), at the present time responses of Excellent are represented as a value of one. In addition to student scores from the above stated sixteen items, course information of class size, percent responding, instructional mode, foundational area, and GEP theme were included in the analysis. No individual course, student, or instructor information was retained in the data set used for analysis.

Statistical and data mining techniques were utilized to extract information regarding association and relationship of individual items to the overall item response. Association of the individual items to the overall item was conducted using Spearman’s correlation (r_s). Decision tree analysis was used to declare rules relating to “Excellent” or “Poor” overall instructor ratings based on the other variables available in this study. Decision tree analysis identified important individual items related to the overall instructor evaluation score.

CHAPTER 4
ANALYSIS OF DATA

Chapter 4 presents summary results, statistical test conclusions, and modeling results from the compiled data set used for analysis. The final data set included information from the student evaluation form and enrollment information for all GEP courses taught between August 2002 and May 2009 at a large public university located in a metropolitan area in the southeast U.S. This research contributes unique information to the topic of instructor evaluations by specifically focusing on GEP courses and the influence of program construction (e.g., class size, foundational area, instructional mode, and theme) on student perception. The purpose of this analysis was to: (a) determine the relationship of individual items to the final or overall rating item; (b) examine variables of class size, percent responding, instructional mode, GEP foundational area, and GEP theme in relation to the overall rating item; and (c) find rules relating individual items to the determination of an “Excellent” overall instructor evaluation or a “Poor” overall instructor evaluation score.

Chapter 4 starts with data screening and data set construction. Individual variable summary information is presented in the third section followed by results for each of the four research questions.

Data Screening

Prior to analysis, the information was checked for inconsistencies and for items that needed to be modified or removed before answering the research questions. Student evaluation information was entered in an electronic data format using a data scanner, so coding problems were minimal. Raw evaluation data provided by Office of Instructional Research initially contained lab sections that were subsequently removed. Using SQL coding statements, values for items 1 to 16 were checked to make sure all coded responses were valid data values (1 to 5). There were a total of 3,490 responses coded as a value of nine, which is not a valid response code, these values were changed to indicate a missing response; these changes amounted to a very small percentage (0.07%) of the total number of values. If students did not respond to the overall rating item (Q16), the entire observation (Q1 to Q16) was deleted. A missing value for the overall rating occurred in 17,249 responses which is approximately 5.5% of all responses. Additionally, if a response had all items 1 through 15 as missing, the response was removed; there were 38 (0.01%) observations removed for this reason.

When the number of students responding was higher than the enrollment figure, then percentage enrollment was set to missing; this procedure was initiated to keep all sections in the analysis, while maintaining proportion responding within the theoretical limits of zero to one. This inconsistency occurred 61 times (0.75%) out of the 8,169 course sections provided in the data set.

Internal consistency problems can occur when students do not take their task of form completion seriously. Responses with an overall evaluation score of “Excellent” or “Poor” had the remaining 15 item responses reviewed in order to identify cases where the responses were clearly not credible. If the response to all individual items was “Poor” or “Excellent” coupled with an “Excellent” or “Poor” (respectively) overall evaluation score, then the observation was removed from analysis. There were only 17 (0.005%) observations out of the total of 294,709 observations that needed to be removed for reasons of credibility.

Data Set Construction

For each academic year, there are three semesters (fall, spring, and summer) with each semester contained in a separate data set. A total of twenty data files, one for each semester (fall

2002 through spring 2009), were supplied from Office of Institutional Research containing student response values for each GEP course taught during that semester. A header record identifying the college, department, instructor, course, and section, separated each group of student responses. During the merging process, each data file had an additional variable for year/semester added.

In addition to the 16 items on the student form, additional variables were created, using SAS coding statements, to identify foundational area, and GEP theme. Another set of files supplied enrollment values and instructional mode categories. Instructional mode refers to how the students interacted with the instructor. The different types for GEP classes were: face-to-face, email enhanced, video feed, ITV feed, face-to-face with video feed, reduced seat time, tech classroom, two-way television, video stream, or web-based. Instructional mode information was collapsed into face-to-face or other due to the small number of courses, 991 (12.29%) using other modes of instructional delivery.

Year/semester information was used to create a binary variable representing GEP theme (yes, no). The first four academic years in the data set (2002-2003, 2003-2004, 2004-2005, 2005-2006) were years prior to the instituted theme of global climate change. The remaining terms of fall 2006 through spring 2009 were with the theme in place. This division resulted in a 55.48% (*n*=163,498) and 44.52% (*n*=131,194) split of the responses respectively for prior to and after the GEP theme was implemented.

Percent responding was calculated based on the number of students responding divided by the number of students officially enrolled in the course. Overall response rate (enrolled students = 480,684) for the semesters used in this study (fall 2002 to spring 2009) was 61.31%.

At the conclusion of the combining process there was one master file constructed, which was used for analysis. The final data set used for analysis contained sixteen student response scores (Q1-Q16), response number, class size, term, foundational area (communication, cultural/historical, mathematics, social, or science), GEP theme (yes/no), instructional mode (face-to-face or other), and percent responding (calculated value). All identifying information such as department, course, section, and instructor was removed from the analysis file. The final data set contained 23 variables, 8,065 course sections, and 294,692 student responses.

Individual Variable Summary Information

Percent responding and enrollment data were available for each course and summary information at the course level (enrolled and percent responding) was assigned to each student response. Using all individual responses (*n* = 294,692), the correlation between class size and percent responding showed a negative relationship (*r* = -0.52). Figure 1 is a scatterplot of the number of students enrolled and percentage of students responding to the course evaluation (*n* = 8065 courses).

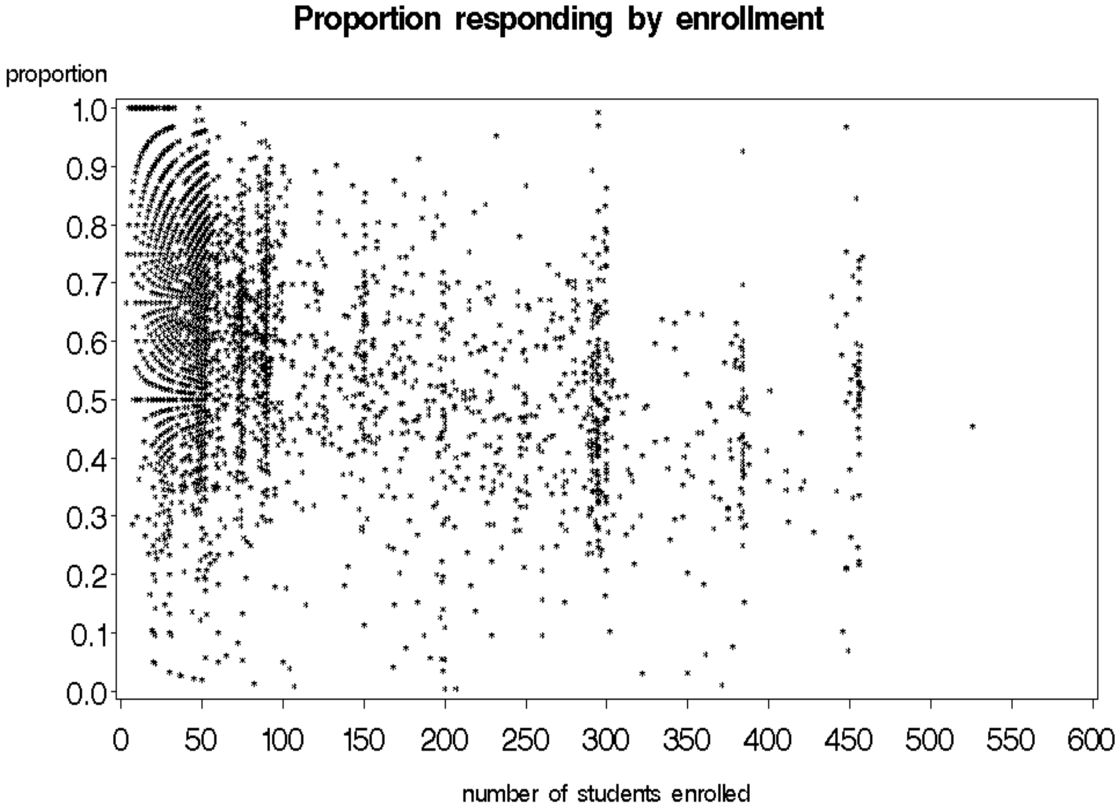


Figure 1. Plot of response proportion by class size enrollment.

Table 2 shows the average class size, number of students enrolled, and percent of students responding to the evaluation form for each foundational area. The average class size was largest for science courses (137.41) and smallest for communications courses at 27.12 students per course. Class sizes ranged from 1 student to a maximum of 526 students. Class sizes of one were typically for classes with the major instructional mode being an alternate method as opposed to face-to-face. Percent of students responding ranged from less than 1% to 100%. The average percentage responding was highest for communications (77.43%) and was lowest for mathematics foundational area courses at 47.55%.

Table 2

Mean class size, number enrolled, and percent responding by foundational area

	Mean	Students	
Foundational Area	Class size	Enrolled	Percent Responding
Communications	27.12	102,702	77.43%
Historical and Cultural	63.61	133,525	64.15%
Mathematics	93.95	57,687	47.55%
Science	137.41	76,127	50.07%
Social	124.32	110,643	53.63%

Note. Mean class size = 60.5 students, Mean percent responding = 61.31%

Table 3 shows the number of courses, average percent responding, and average class size by instructional mode and GEP theme variable. The average percent responding was larger for face-to-face (71.14%) compared to other instructional modes (53.48%). Average class sizes (Table3) for each instructional mode and prior to or with the GEP theme in place range from 32 to 39 students.

Table 3

Number of courses, mean class size, and percent responding by mode and GEP theme

	Instructional Mode		GEP theme	
	Face-to-face	Other	No	Yes
Number of sections (percentage)	7074 (87.71%)	991 (12.29%)	4703 (58.31%)	3362 (41.69%)
Average percent responding	71.14%	53.48%	69.16%	69.24%
Mean class size	37.18	32.20	35.78	39.06

Based on all of the information available, 43.82% of students regarded their instructor as “Excellent” with only 2.26% indicating their overall rating belonged in the “Poor” category. Table 4 presents percentages for each of the five categorical responses by foundational area and summarized for the entire data set. The results showed that courses in the mathematics foundational area had the lowest proportion of “Excellent” overall scores (35.27%) of any foundational area. Communications and historical and cultural foundation areas had the highest proportion of “Excellent” overall scores. Response categories within each foundational area followed a similar decreasing pattern when viewed from “Very good” to “Poor” categories.

Table 4

Percentages for Overall rating (Q16) category by Foundational Areas

	Foundational Area					Summary
Overall rating	Communi-cations	Historical & Cultural	Math	Science	Social	Total
Excellent	46.42	46.40	35.27	39.46	43.35	43.82
Very Good	28.88	29.28	28.90	27.27	30.37	29.10
Good	16.68	16.88	23.33	21.27	18.59	18.36
Fair	5.87	5.48	9.35	8.75	5.89	6.46
Poor	2.15	1.96	3.14	3.26	1.80	2.26

Note. Each column sums to 100%, *n* = 294,692

Ranking of foundational areas from high to low in terms of percentage of "Excellent" overall scores was: communications, historical and cultural, social, science, and mathematics. Interestingly, this is the same ranking as seen previously in Table 2. This indicates that higher percentage responding foundational areas also have a higher percent of "Excellent" overall scores given by students.

Table 5 presents the percentages for each of the five categorical responses by instructional mode. Face-to-face courses accounted for 87.71% of all GEP courses and 44.12% of student responses designated the instructor as "Excellent." All other instructional modes accounted for the remaining 12.29% with 41.30% of student responses designating an

"Excellent" overall evaluation score. Response categories (Excellent to Poor) followed a similar decreasing pattern across both instructional type course groups.

Table 5

Percentages for Overall rating (Q16) category as a function of instructional type

	Overall Rating Response				
Instructional Mode	Excellent	Very good	Good	Fair	Poor
Face-to-face (7074 course sections)	44.12	28.96	18.26	6.45	2.20
All others (991 course sections)	41.30	30.28	19.14	6.57	2.71

Note. Percentage of sections Face-to-face (87.71%), other modes (12.29%).

Table 6 shows percentages for each of the five categorical responses by the variable indicating whether or not the GEP theme of global climate change was initiated. The percentage of "Excellent" overall scores was slightly higher (45.78%) for courses with the GEP theme in place compared to 42.24% prior to the current GEP theme. Response categories (e.g., Excellent to Poor) followed a similar decreasing pattern across both groups.

Table 6

Percentages for Overall rating (Q16) during semester with and without GEP theme

	Overall Rating Response				
GEP theme	Excellent	Very good	Good	Fair	Poor
Not in place (4703 course sections)	42.24	29.51	19.01	6.78	2.45
Initiated (3362 course sections)	45.78	28.59	17.55	6.06	2.03

Note. Twelve semesters prior to theme and eight semesters after current theme started.

Table 7 illustrates the percentage of responses for each category for the fifteen individual items on the student evaluation form. The highest percentage of “Excellent” scores (46.88%) was for item 13 which asked about instructors’ respect and concern for students. The lowest percentage for “Excellent” was for item 8 (26.47%) which solicited a response to rating the textbook and supplemental materials used in the course. A similar decreasing pattern in the percentages for each row was evident.

Table 7

Percent of responses for each item and category

	Response					
Item	Excellent	Very good	Good	Fair	Poor	Missing
1	30.16	31.54	25.87	9.06	2.54	0.83
2	39.14	31.05	20.56	6.89	1.85	0.52
3	37.42	31.02	20.93	7.42	2.57	0.65
4	39.59	31.10	19.90	6.69	2.24	0.48
5	42.44	31.41	19.39	4.89	1.23	0.64
6	32.68	31.91	24.85	7.84	2.17	0.55
7	29.41	29.72	26.73	10.19	3.06	0.89
8	26.47	27.82	28.82	12.13	3.92	0.83
9	35.93	32.12	22.85	6.46	2.09	0.56
10	38.31	29.76	20.77	7.50	3.10	0.56
11	34.34	31.03	24.49	7.32	2.11	0.71
12	35.71	28.12	24.31	7.79	2.44	1.62
13	46.88	27.05	17.66	5.38	2.42	0.61
14	38.05	27.00	21.17	8.78	4.36	0.64
15	35.78	30.56	23.72	6.71	2.44	0.79

Note. Number of responses = 294,692

Research Question 1

What items on the student evaluation form are related to the "Overall rating" score provided by students? Spearman correlation coefficient (r_s) was computed to measure association because of the ordinal composition of the items contained on the evaluation form. Spearman's correlation (r_s) uses ranks to calculate a measure of relationship between two variables from $-1 \leq r_s \leq +1$ (McClave & Sincich, 2006). Spearman values closer to either -1 or +1 indicate strong association. Association of items one through fifteen with the overall score were all statistically significant ($p<0.001$). Very small significance levels are not unusual when the number of observations is extremely large.

The formula for the Spearman's rank correlation value is given below. In the following formula n refers to the number of pairs of observations, u_i refers to the rank of the ith observation in sample 1, v_i refers to the rank of the ith observation in sample 2.

$$r_s = \frac{\sum (u_i - \overline{u})(v_i - \overline{v})}{\sqrt{\sum (u_i - \overline{u})^2 \sum (v_i - \overline{v})^2}} \qquad (2)$$

Table 8 shows Spearman correlation values for each of the individual items and the overall score item. These results show that the highest correlations with the overall rating (item 16) was for item 10, communication of ideas, and item 15 which asks about facilitation of learning. The lowest correlation with the overall rating (item 16) was for item 8 which asks students to score the text and supplemental materials used in the course. Logically, this item has the least association with the overall score given to the instructor as compared to other items on the evaluation form. Appendix C contains a table, in matrix form, of Spearman correlation values for items one through fifteen.

Table 8

Spearman correlation values for individual items with overall evaluation item (ans16)

Description	Variable	Spearman (r_s)
Feedback concerning performance	Q1	0.69
Instructor’s interest in your learning	Q2	0.77
Use of class time	Q3	0.73
Instructor’s overall organization	Q4	0.76
Continuity for one class to the next	Q5	0.72
Pace of the course	Q6	0.72
Instructor’s assessment of your progress	Q7	0.74
Texts and other material	Q8	0.58
Description of course objectives and assignments	Q9	0.74
Communication of ideas and information	Q10	0.81
Expression of expectations for performance	Q11	0.77
Availability to assist students in or outside of class	Q12	0.69
Respect and concern for students	Q13	0.77
Stimulation of interest in the course	Q14	0.78
Facilitation of learning	Q15	0.80

Research Question 2

Is instructional mode, class size, GEP area, or incorporating a GEP theme related to the "Overall rating" (item 16) score provided by students? Data mining modeling techniques were used to understand the relationship of these course variables with the overall score value from the student evaluation form. The rating categories for the variable of interest were: Excellent, Very Good, Good, Fair, and Poor. Item one to fifteen (Q1 to Q15), instructional mode, foundational area, percent responding, and GEP theme were the independent variables used to find splits to create homogenous groups in relation to the overall evaluation score.

Class size was removed from the final list of variables used in modeling due to anomalies for a very small range of class sizes. The mean class size for this study was 60.04 students with a standard deviation of 58.44 students. Investigation of these six courses showing differences in overall scores did not reveal any reasons for removing the associated student responses. The courses did have typical values for percent responding and yet the overall evaluation score (item 16) distribution was slightly different from the general group of student responses. Because of these anomalies, percent responding instead of class size was used to understand if classroom size factors might be related to student perception.

Enterprise Miner™ was used for the decision tree analysis (SAS Institute, 2008). Decision tree analysis was selected to model this data because of the data composition and data structure. Responses with missing values were included in data mining analysis and were not imputed prior to analysis. Inclusion of missing values is an advantage of data mining tree modeling tools. The structure of student evaluation response data is unknown, but probably is not linear, which makes decision tree analysis an excellent option for partitioning the data. When the underlying data structure is unknown, it is difficult to use statistical techniques, such as regression analysis, which may require certain relationships among variables. Decision trees can easily partition very large data sets containing both continuous and categorical data. Decision tree results can be stated as easily interpreted rules for understanding the overall evaluation score. Each rule has information in the form of odds ratios and misclassification rates that can guide interpretation.

Research question two specifically asked about the relationship of variables external to the student evaluation form. The first analysis generated included all variables from the student evaluation form and variables for instructional mode, foundational area, percent responding, and GEP theme. However this first decision tree model did not use any of the variables external to the student evaluation form as important variables for the overall evaluation score. This indicated there were no significant differences in how students arrive at an overall value related to differences in instructional mode, foundational area, percent responding, or GEP theme. The only variables showing a relationship with the overall instructor score were individual items from the student evaluation form.

The second analysis included only variables external to the form (foundational area, instructional mode, percent responding, GEP) in the modeling of overall instructor score. This model was not examined in detail due to the insignificance of the external variables when considered as part of the complete list of variables included in this study.

In this model, including only course variables (percent responding, foundational area, instructional mode, and GEP theme) the most important variable was percent responding. As the percent responding increased above 66.55% the proportion of “Excellent” overall scores increased from 43.8% to 48.0%. The next split pertained to the group of responses with percent responding less than 66.55% which revealed the GEP theme had a higher percentage “Excellent” than the GEP classes prior to the theme. The third split revealed that foundational area further divided the percentage of "Excellent" scores with mathematics and science foundational areas having the lower percentage of "Excellent" when compared to the other foundational areas.

Figure 2 shows the software nodes constructed to determine the decision tree model using all independent variables (model 1) and a separate analysis using only four variables external to the student evaluation form (model 2). The left most icon represents the data set used for analysis. The data partition icon divides the responses into two different groups to be used in the decision tree analysis. The mulitplot icon allows for data exploration using graphs and plots. The final two icons (on the right) are the decision tree analysis models. Figure 3 shows the results of the decision tree analysis using only four variables external to the student evaluation form. The details of model two were described previously. Model one will be used to answer the remaining research questions (questions three and four) and will be described in detail in the following section.

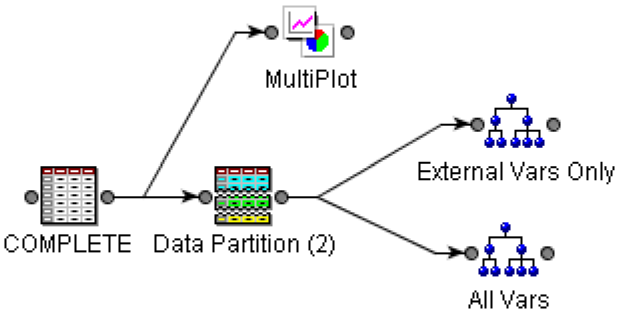
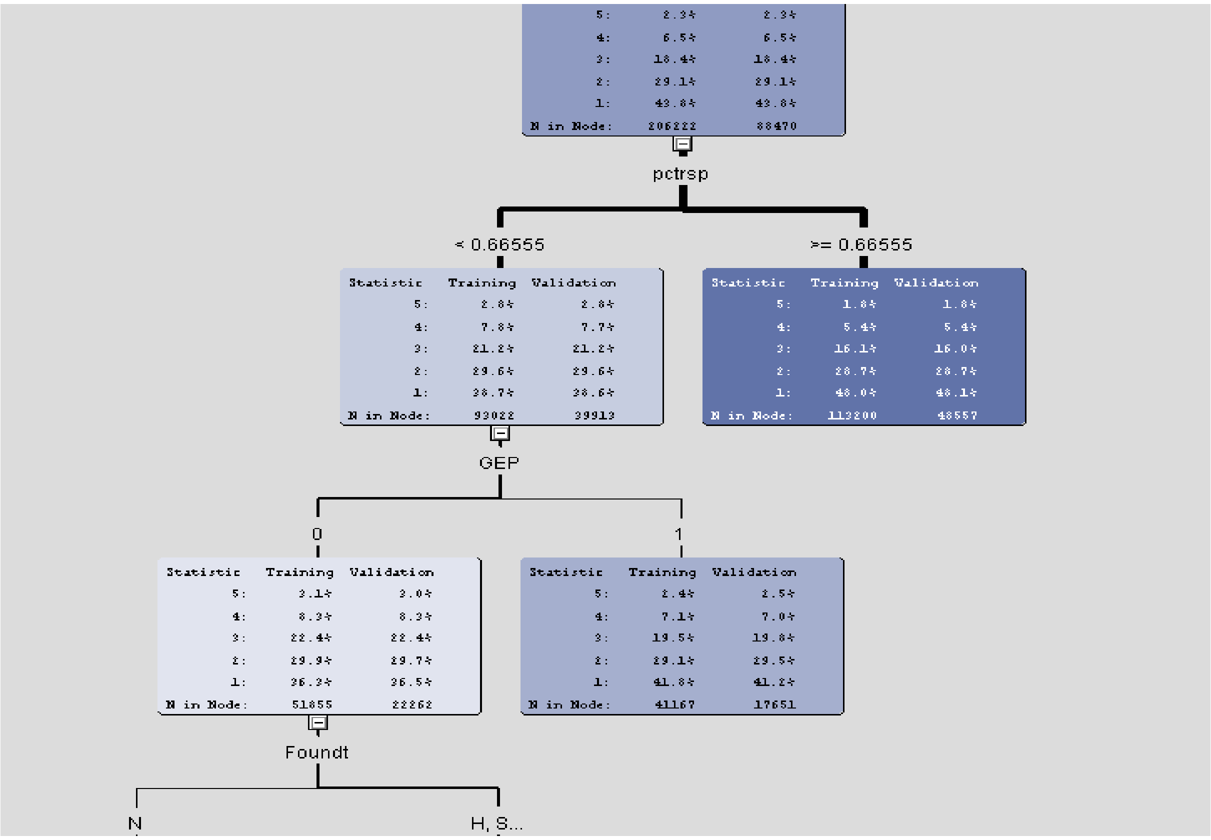


Figure 2. SAS® Enterprise Miner™ screen shot for decision tree analysis.



5:	2.3%	2.3%
4:	6.5%	6.5%
3:	18.4%	18.4%
2:	29.1%	29.1%
1:	43.8%	43.8%
N in Node:	206222	88470

Statistic	Training	Validation
5:	2.8%	2.8%
4:	7.8%	7.7%
3:	21.2%	21.2%
2:	29.6%	29.6%
1:	38.7%	38.6%
N in Node:	93022	39913

Statistic	Training	Validation
5:	1.8%	1.8%
4:	5.4%	5.4%
3:	16.1%	16.0%
2:	28.7%	28.7%
1:	48.0%	48.1%
N in Node:	113200	48557

Statistic	Training	Validation
5:	3.1%	3.0%
4:	8.3%	8.3%
3:	22.4%	22.4%
2:	29.9%	29.7%
1:	36.3%	36.5%
N in Node:	51855	22262

Statistic	Training	Validation
5:	2.4%	2.5%
4:	7.1%	7.0%
3:	19.5%	19.8%
2:	29.1%	29.5%
1:	41.8%	41.2%
N in Node:	41167	17651

Figure 3. SAS® Enterprise Miner™ decision tree results for external variables only.

Research Questions 3 and 4

Decision tree analysis was used to answer research questions regarding what rules can be identified to understand the determination of an "Excellent" overall rating or a "Poor" overall rating on the student evaluation form. The target variable was the overall instructor score, item 16 on the form. Item 1 to item 15, instructional mode, foundational area, and GEP theme were the independent variables used to find splits to create homogenous groups. Rating categories for all items on the student evaluation form were: Excellent, Very Good, Good, Fair, and Poor.

The data set was split 70% and 30% into training and validation data subsets prior to modeling. The training data set generates the initial model which is typically over specified and thus is modeling more error than would be preferred. Partitioning the data set allows for the tree structure to be grown larger than necessary and then pruned back to the size specified as optimum. Optimum size of the final tree model is specified by adjusting software settings within the decision tree analysis icon. Because the validation data is independent of the original model structure, it gives an honest estimate of the best tree size. In other words, the initial model is applied to new responses and fine tuned based using the information in the validation data set.

Decision tree models are built by segmenting the observations into smaller and smaller groups. Model results are in the form of multiple if-then statements. For each node or group the probability of correct classification is supplied via software options; higher probabilities indicating homogeneity of observations in regard to the target variable. In order to determine the overall tree model results, a global misclassification rate was used as the criterion of model homogeneity. The final model had a misclassification rate of 25.5% for the training data set and 26.0% for the validation data set. Interpretation of the misclassification rate is situation specific (Wang, 2007).

The top three rules for “Excellent” and “Poor” based on probability will be outlined in the results. The number of rules was selected by considering the percentage of responses and the number of responses that conform to that particular rule. Three rules with the highest percentage of response conformity were selected to be discussed. All resulting rules based their classifications on some combination of the following items: communication of ideas and information; facilitation of learning; respect and concern for students; instructor’s overall organization of the course; instructor’s assessment of your progress in the course; instructor’s interest in your learning; stimulation of interest in the course. Figure 4 shows the tree diagram branching display; visually describing the model structure.

Excellent overall score rules

The final tree model (figure 4) resulted in rules determining the probability of obtaining an overall instructor score of Excellent when certain conditions held within the complete data set. The top three rules have the highest probability based on scores from other items on the student evaluation form. All the rules for “Excellent” used only items on the evaluation form and eliminated foundational area, class size, percent responding, instructional mode, and GEP theme information. The top three rules for predicting an “Excellent” overall evaluation rating are summarized in Table 9. Rules for an “Excellent” overall score based their classifications on some combination of the following items: *communication of ideas and information; facilitation of learning; respect and concern for students; instructor’s overall organization of the course; instructor’s assessment of your progress in the course*. Of the 109,759 responses that conformed to one of the rules, the largest percentage (81.6%) belonged to rule one, followed by rule two (10.9%) and rule three (7.5%).

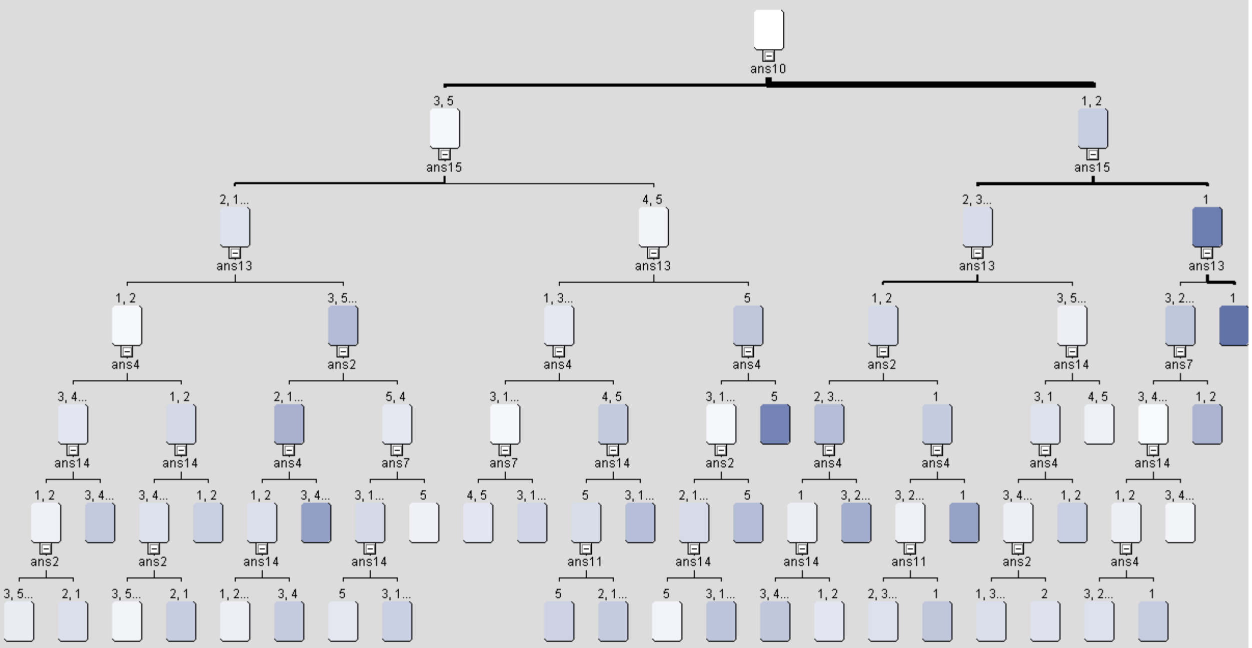


Figure 4. Decision Tree results for prediction of overall instructor score.

Rule one produced the highest rating for an “Excellent” overall score with an associated probability of 0.94; indicating that if the conditions for those variables in the rule hold true, then for those responses 94% of the responses would have received an “Excellent” overall score. Rule one states that when *communication of ideas and information* was “Excellent” or “Very good” and *facilitation of learning* and *respect and concern for students* items were rated as “Excellent,” then the result was generally an “Excellent” overall score. In the data set as a whole, the percentage of “Excellent“ overall scores was 43.82% thus making the odds ratio of rule one equal to 2.15; odds ratios indicate the odds of getting an overall score of “Excellent” if you receive the above scores represented in the rule. Odds ratios are calculated by dividing the specific rule probability for the target variable by the probability for the entire data set of having that same target variable value. For example, odds of 2.15 result from using the rule probability (0.94) divided by the overall probability of receiving an “Excellent” overall score (0.4382).

Rule two produced a rating for an "Excellent" overall score with an associated probability of 0.78; indicating that if the conditions for those variables in the rule hold true, then for those responses 78% of the responses would have received an "Excellent" overall score. Rule two combines information from the following five individual items: *communication of ideas and information* as "Excellent" or "Very good"; *facilitation of learning* as "Very good," "Good," or "Fair"; *respect and concern for students* as "Excellent" or "Very good"; *instructor's interest in your learning* as "Excellent"; *instructor's overall organization of the course* as "Excellent". In the data set as a whole, the percentage of "Excellent" overall scores was 43.82% thus making the odds ratio of rule one equal to 1.78; odds ratios indicate the odds of getting an overall score of "Excellent" if you receive the above scores represented in the rule.

Rule three has an associated probability of 0.71; indicating that if the conditions for those variables in the rule hold true, then for those responses 71% of the responses would have received an "Excellent" overall score. Rule three contains four items, three of which were in Rule one. Individual items for Rule 3 consisted of: *communication of ideas and information* as "Excellent' or "Very good"; *facilitation of learning* as "Excellent"; r*espect and concern for students* as "Very good" or "Good"; *instructor's assessment of your progress* as "Excellent" or "Very good". The odds ratio for rule two was 1.62; meaning these instructors are more than one and a half times as likely to receive an "Excellent" overall rating as one drawn at random.

When "Excellent" and "Very good" overall scores were combined, all three rules (one, two, and three) had predictive probabilities of 99%. In other words, if the rule identified a student response, 99% of these responses had an overall score of "Excellent" or "Very good."

Table 9

Decision Rules that Lead to an Overall Instructor Rating of "Excellent"

Question	Rating					Excellent proportion (*proportion for Excellent and Very Good*)
	E	VG	G	F	P	
Rule 1 (*n* = 89,592)						
Communication of Ideas and Information	●	●				.94
Facilitation of Learning	●					(.99)
Respect and Concern for Students	●					
Rule 2 (*n* = 11,947)						
Communication of Ideas and Information	●	●				
Facilitation of Learning		●	●	●		.78
Respect and Concern for Students	●	●				(.99)
Interest in Student Learning	●					
Overall Organization of the Course	●					
Rule 3 (*n* = 8,220)						
Communication of Ideas and Information	●	●				
Facilitation of Learning	●					.71
Respect and Concern for Students		●	●			(.99)
Assessment of Progress	●	●				

Poor overall score rules

The final tree model resulted in rules determining the probability of obtaining an overall instructor score of Poor when certain conditions held within the complete data set. The top three rules have the highest probability based on scores from other items on the student evaluation form. All the rules for "Poor" used only items on the form and eliminated foundational area, class size, percent responding, instructional mode, and GEP theme information. The top three rules for predicting a "Poor" overall evaluation rating are summarized in Table 10. Rules for a "Poor" overall score based their classifications on some combination of the following items: *communication of ideas and information; facilitation of learning; stimulation of interest in the course; respect and concern for students; instructor's overall organization of the course; instructor's interest in your learning; expression of expectations for performance*. Of the 3,821 responses that conformed to one of the rules, the largest percentage (56.1%) belonged to rule one, followed by rule two (27.7%), and rule three (16.2%).

Rule four produced the highest rating for a "Poor" overall score with an associated probability of 0.89; indicating that if the conditions for those variables in the rule hold true, then for those responses 89% of the responses would have received a "Poor" overall score. Rule four reveals that when *communication of ideas and information* was "Good," "Fair," or "Poor" and *facilitation of learning* was "Fair" or "Poor" along with *respect and concern for student* and *overall organization* items are rated as "Poor," then the result was generally an "Poor" overall score. In the data set as a whole, the percentage of "Poor" overall scores was 2.26% thus making the odds ratio of rule four equal to 39.2; odds ratios indicate the odds of getting an overall score of "Poor" if you receive the above scores represented in the rule. Because the percentage of

"Poor" overall instructors was so low (2.26%), the odds ratio may overestimate the magnitude of this likelihood.

Rule five produced a probability of 0.68 for a "Poor" overall score; indicating that if the conditions for those variables in the rule hold true, then for those responses 68% of the responses would have received a "Poor" overall score. Rule five combines information from the following five individual items: c*ommunication of ideas and information* was "Good," "Fair," or "Poor"; *facilitation of learning* was "Fair" or "Poor"; *respect and concern for students* was "Poor"; *overall organization* was "Good" or "Fair"; *instructor's interest in your learning* was "Poor". The odds ratio for rule five was 29.96; meaning these instructors are much more likely to receive a "Poor" overall rating as one drawn at random.

Rule six has an associated probability of 0.60; indicating that if the conditions for those variables in the rule hold true, then for those responses 60% of the responses would have received a "Poor" overall score. Rule six contains six items, four of which were in Rule four. Individual items for Rule six were: *communication of ideas and information* at "Good," "Fair," or "Poor"; *facilitation of learning* at "Fair" or "Poor"; *respect and concern for students* at "Excellent," "Very good," "Good," or "Fair"; *overall organization* at "Fair" or "Poor"; *stimulation of interest in the course* at "Poor"; *expression of expectations for performance* at "Poor". The odds ratio for rule six was 26.55; meaning these instructors are more than twenty-six times as likely to receive a "Poor" overall rating as one drawn at random.

When "Poor" and "Fair" overall scores were combined, all three rules (four, five, and six) had predictive probabilities of 99%. In other words, if the rule identified a student response, 99% of these responses had an overall score of "Poor" or "Fair."

Table 10

Decision Rules that Lead to an Overall Instructor Rating of "Poor"

	Rating					Poor proportion (*proportion for Poor and Fair*)
Question	E	VG	G	F	P	
Rule 4 (*n* = 2,143)						
Communication of Ideas and Information			●	●	●	.89
Facilitation of Learning				●	●	(.99)
Respect and Concern for Students					●	
Overall Organization of the Course					●	
Rule 5 (*n* = 1,057)						
Communication of Ideas and Information			●	●	●	
Facilitation of Learning				●	●	.68
Respect and Concern for Students					●	(.99)
Overall Organization of the Course			●	●		
Interest in Student Learning					●	
Rule 6 (*n* = 621)						
Communication of Ideas and Information			●	●	●	
Facilitation of Learning				●	●	.60
Respect and Concern for Students	●	●	●	●		(.99)
Interest in Student Learning				●	●	
Expectations for Student Performance					●	
Stimulation and Interest in the Course					●	

In order to check the validity of the data mining results each of the six rules was applied to each foundational group separately. If the rule conformity percentages are similar across foundational areas, this demonstrates the consistency of the overall model. Table 11 demonstrates the impact of the three “Excellent” rules on each of the GEP foundational areas. The unadjusted percentage for each rating group by foundational level was presented in Table 4. The unadjusted percentage for “Excellent” ranged from a maximum value of 46.42% for Communications foundational area to a low of 35.27% for Mathematics foundational area. Differences in percentage of “Excellent” scores disappear when rule one was applied to each foundational area separately. These adjusted percentages are given in Table 11. Rule one had a combined percentage of 94% for the entire data set and there is virtually no difference for any foundational area. Rule two percentages moderated slightly when considered within each foundational area. Communications produced a high of 79.68% compared to a low of 73.45% from the Science foundational area. Rule three percentages were only slightly moderated, with 74.46% as the high for Science and a low of 67.69% for Communications.

Table 11

Percent of Excellent overall ratings by foundational area adjusted for rules 1-3

	Adjusted					
Foundational Area (unadjusted percentage)	*n*	Rule 1 (94%)	*n*	Rule 2 (78%)	*n*	Rule 3 (71%)
Communications	25,387	94.39	3,898	79.68	1,996	67.69
Historical and Cultural	28,926	94.03	3,345	77.43	2,646	70.67
Mathematics	6,541	94.08	973	77.49	692	68.06
Science	10,745	94.16	1,439	73.45	979	74.46
Social	17,993	94.43	2,292	76.66	1,907	73.10

Table 12 demonstrates the impact of the three “Poor” rules on each of the GEP foundational areas. The unadjusted percentage for each rating group by foundational level was presented in Table 2. The unadjusted percentage for “Poor” ranged from a high value of 3.26% for Science to a low of 1.80% for Social foundational area. The differences in percentage of “Poor” scores disappear when rule one was applied to each foundational area separately. Rule one had a percentage of 89% for the entire data set and there is very little difference for any foundational area. Rule two percentages changed from the 68% average when considered within each foundational area. Communications produced a high of 76.00% compared to a low of 62.78% from the Social foundational area. Rule three percentages were different by foundational area, with 69.43% as the high for Communications and a low of 52.83% for Social.

Table 12

Percent of Poor overall ratings by foundational area adjusted for rules 4-6

Foundational Area (unadjusted percentage)	Adjusted					
	n	Rule 4 (89%)	*n*	Rule 5 (68%)	*n*	Rule 6 (60%)
Communications	555	88.29	200	76.00	157	69.43
Historical and Cultural	565	88.14	258	69.38	128	66.41
Mathematics	267	90.64	183	67.76	107	57.94
Science	421	89.07	236	72.03	123	69.11
Social	335	91.64	180	62.78	106	52.83

Summary

In this study, student evaluation scores from all GEP courses in consecutive academic years 2002-2003 ending spring 2009 were used as the data set of interest. The final data set contained 23 variables, 8,065 course sections, and 294,692 student responses. Student evaluation information consisted of sixteen items measured on an ordinal scale with values from 1 to 5 representing categories of Excellent, Very good, Good, Fair, and Poor. In addition to student scores from the above stated sixteen items, course information of class size, percent responding, instructional mode, foundational area, and GEP theme were included in the analysis.

Preliminary variable investigation revealed consistent trends in percentages of response for the overall instructor score when separated by instructional mode, foundational area, class size or GEP theme. Percent responding calculated as number of students responding to the evaluation form divided by number of students enrolled in the course was used to understand whether class size differences were present in the data set; number of students enrolled was eliminated.

Based on all the student information available, 43.82% of students regarded their instructor as “Excellent” with only 2.26% indicating their overall rating belonged in the “Poor” category. Only 12.29% of course sections were offered via instructional mode other than face-to-face format. GEP theme of “global climate change” was instituted in fall 2006 with 45.78% of students indicating “Excellent” for the overall instructor rating as compared to 42.24% of students indicating “Excellent” prior to the GEP theme in place.

Foundational area had consistent trends within each area for percentages of overall rating scores (i.e., Excellent to Poor), but differences between foundational areas, with mathematics and science having the lowest percentage of “Excellent” scores. Other preliminary results revealed a negative correlation between class size and the proportion of students responding to the evaluation.

Research question one found that all items on the student evaluation form are related to the “Overall rating” score provided by students. This association was measured using Spearman correlation values.

Research question two revealed that course variables of percent responding, instructional mode, foundational area, and GEP theme were not selected by the decision tree model as important variables for the overall evaluation score when the fifteen individual items on the form were also included in the modeling process. This indicated there were no significant differences in how students arrive at an overall value related to differences in the defined course variables.

Decision tree analysis was used to answer research questions regarding what if-then rules can be identified to understand the determination of an “Excellent” overall rating or a “Poor” overall rating on the student evaluation form. Decision tree modeling results in if-then type rules that can be used to categorize the responses. All the if-then rules used only items on the evaluation form and eliminated foundational area, class size, percent responding, instructional mode, and GEP theme information.

Rules for an “Excellent” overall score based their classifications on some combination of the following items: *communication of ideas and information; facilitation of learning; respect and concern for students; instructor’s overall organization of the course; instructor’s assessment of your progress in the course*. Proportion of student responses conforming to the top three rules for “Excellent” overall evaluation ranged from .94 to .71.

Rules for a “Poor” overall score based their classifications on some combination of the following items: *communication of ideas and information; facilitation of learning; stimulation of interest in the course; respect and concern for students; instructor’s overall organization of the course; instructor’s interest in your learning; expression of expectations for performance*. Proportion of student responses conforming to the top three rules for “Poor” overall evaluation ranged from .89 to .60.

When the two highest overall classifications (Excellent and Very good) or the two lowest overall classifications (Poor and Fair) were combined, the rules (one through six) had predictive probabilities of 99%.

CHAPTER 5
CONCLUSIONS

Introduction

Chapter 5 provides discussion of the data analysis from Chapter 4 and begins with the purpose of this research, followed by preliminary analysis. Each research question is subsequently summarized with important questionnaire items addressed and synthesis of the results. The chapter concludes with implications of the results and future research avenues.

Purpose

The purpose of this study was to determine, for GEP courses, what instructor qualities can be identified as critical factors for scores obtained via student evaluation forms. These factors were compiled using individual items on the student evaluation form or course factors that are predictive of the overall instructor rating value. The study examined what instructor/course qualities were related to a high overall evaluation or a low overall evaluation value thus placing an overall evaluation in the context of actionable items for improvement. Additional information on course year, instructional mode, GEP foundational area, and class size was examined in relation to overall scores given by students. This research contributes unique information to the topic of instructor evaluations by specifically focusing on GEP courses and the influence of program construction (i.e., class size, theme, foundational area) on student perception.

The purpose of the analysis was to: (a) determine the relationship of individual items to the final or overall rating item; (b) examine whether or not course characteristics of class size, instructional mode, GEP foundational area, and GEP theme were related to the overall rating item; and (c) find rules relating individual evaluation items or course characteristics to the determination of an “Excellent” overall evaluation or a “Poor” overall evaluation score.

Preliminary Analysis

Overall response rate was 61.31%, which was generated from 294,692 student responses over the twenty semesters of interest. Response rates varied by instructional mode, class size, and foundational area but did not differ for GEP theme variable.

Out of the 8,065 sections used in this study, 7,074 or 87.71% were conducted as face-to-face instruction while the remaining 991 or 12.29% used an alternative instructional mode. Course delivery mode was important because of the relationship with evaluation procedures. For this time period (2002-2009) all face-to-face courses used the paper version of the student evaluation form. Other instructional modes (email enhanced, video feed, ITV feed, face-to-face with video feed, reduced seat time, tech classroom, two-way television, video stream and web-based) may have had the option of using paper or an electronic version of the evaluation form. Although there is no information on how many observations came from electronic completion, due to the high percentage of face-to-face classes, the majority of responses in this study most likely came from in-class completion of the student evaluation form. Alternative instructional modes had a response rate of 53.48% compared to face-to-face courses with a response rate of 71.14%.

Goyder (2009) in a study at the University of Washington recorded response rates for electronic evaluations at 27% compared to 64% for the paper version. Recent studies reviewed by Goyder (2009) typically show higher response rates for in-class completion, approximately 50-60%, compared to electronic completion of an identical form which had response rates of approximately 30-40%. Response rates for this study were 71.14% for face-to-face courses compared to 53.48% for all other instructional modes combined. This study agrees with the majority of current research showing a lower response rate for electronic student evaluations.

Class size had a strong negative relationship with percent responding (*r* = -0.52). This means that as enrollment size increased the percent of students completing the evaluation form typically decreased. Response rates did vary considerably by foundational area. The results showed that mathematics courses had the lowest percentage response at 47.55% compared to a high of 77.43% for communications. Response rate was examined prior to investigating study research questions.

Analysis

Of primary interest was the overall evaluation score a student assigns to the instructor for that particular course and how this score relates to other items on the form and course variables of instructional mode, GEP theme, percent responding, and foundational area. This item "Overall assessment of instructor" was the final item in a list of sixteen items which could be answered as Excellent, Very good, Good, Fair, or Poor. Classification and regression tree or decision tree analysis was used to find rules to express important items related to the overall evaluation score. These rules are easily interpreted by multiple stakeholders (e.g., instructors, administrators, students) using probability statements and odds ratios. Overall model performance was assessed via misclassification rates.

Study results showed that approximately 44% of students regarded their instructor as "Excellent" with only 2% indicating their overall instructor belonged in the "Poor" category. When "Excellent" and "Very good" were combined, the percentage climbed to 73%.

Examining the overall rating percentages for foundational areas a difference was evident. Communications (46.42%) and historical and cultural (46.40%) foundation areas had the highest proportion of “Excellent” overall scores while mathematics had the lowest proportion of “Excellent” scores (35.27%). Gravestock and Gregor-Greenleaf (2008) report course discipline does have a measurable impact on evaluation ratings. The current study results agree with Feldman (1978) and Gravestock and Gregor-Greenleaf (2008) showing mathematics and natural science courses typically have lower scores whereas literature and history have higher scores.

Scores given in face-to-face courses had a higher percentage (44.12%) of “Excellent” scores and accounted for 87.71% of all GEP courses taught when compared to all other instructional modes which accounted for the remaining 12.29% of courses and had 41.30% “Excellent” overall evaluation scores. When both “Excellent” and “Very good” overall scores were combined, the comparison of percentages becomes 73.08% for face-to-face courses compared to 71.58% for other instructional modes.

A theme for GEP of “Global climate change” was introduced in 2006. Information from twelve semesters prior to theme introduction was compared to eight semesters with the theme in place. The “Excellent” percentage designation went from 42.24% prior to the GEP theme to 45.78% for the courses with the GEP theme in place. The percentage of “Poor” overall scores also decreased from 2.45% to 2.03%, with the GEP theme in place. Although the change in overall score cannot be directly attributed to the addition of a GEP theme, given other conditions remaining constant, the results point to the conclusion that adding a theme to the GEP program was beneficial. Course numbers, and thus basic content, have remained the same except for the addition of one new course in biotechnology and genetics and a reconfiguration of a communications course. Unfortunately, limited information is currently available regarding the number of instructors that incorporated examples from the current theme or the extent to which they employ thematic examples in their courses.

Examining individual items on the form and percentage of responses for each of the five categories showed a similar decreasing pattern of responses from “Excellent” to “Poor.” The highest percentage of “Excellent” scores (46.88%) was for the item which asked about instructors’ respect and concern for students. The *respect and concern* item showed up in all the “Excellent” and “Poor” rules.

Research Question 1

What items on the student evaluation form are related to the “Overall rating” (item 16) score provided by students? The individual items on the student evaluation form were all related to the overall evaluation item score. Spearman’s correlation coefficient was used to measure association because of the ordinal measurement scale of the items contained on the evaluation form. The range for Spearman is $-1 \leq r_s \leq +1$ (McClave & Sincich, 2006). The items with the highest Spearman values were for *facilitation of learning* (0.80) and *communication of ideas and information* (0.81). Both of these items were the most important variables in the rules for “Excellent” and “Poor” overall scores and have large effect sizes. The lowest Spearman value of 0.58 was for the item which asks students to score the text and supplemental materials used in the course. Logically this item has the least association with the overall score given to the instructor as compared to other items on the evaluation form.

Research Question 2

Is instructional mode, class size, GEP foundational area, or incorporating a GEP theme related to the “Overall rating” (item 16) score provided by students? None of the examined course variables were selected as significant when the individual form items were included in the modeling process. This indicated there was no significant difference in how students arrive at an overall value of their instructor related to differences in instructional mode, class size, foundational area, percent responding, or GEP theme. In other words, students employed a consistent approach to evaluation regardless of large or small classes, face-to-face or other instructional modes, foundational area, or percent responding differences.

Research Questions 3 and 4

What rules can be identified to understand the determination of an "Excellent" overall rating on the SET form? What rules can be identified to understand the determination of a "Poor" overall rating on the SET form? Data mining modeling techniques were used to understand the relationship of individual item responses and additional course information variables to the overall score. Items one to fifteen (Q1 to Q15), class size, instructional mode, foundational area, and GEP theme were the independent variables used to find splits to create homogenous groups in relation to the overall evaluation score. The results are presented in terms of rules for "Excellent" or "Poor" overall evaluation scores. All rules based their classifications on some combination of the following items: *communication of ideas and information; facilitation of learning; respect and concern for students; instructor's overall organization of the course; instructor's interest in your learning; instructor's assessment of your progress in the course; and stimulation of interest in the course.* A note of caution when interpreting the results, selection of these items reflect the students' perception of communication, facilitation, etc…, not necessarily the instructors' idea of these concepts.

Communication of ideas and information

Students consistently rated *communication of ideas and information* as the most important item on the form in relation to the overall evaluation score. Communication was present in all six rules and suggests that the ability to communicate effectively is essential to being viewed positively by students. Study results agree with Wang et al., (2009) that communication is an important factor in the modeling of overall score. The authors found the communication item to be important in all rules cited. Their study included both undergraduate and graduate course for years 1996-2001. Wang et al., (2009) state that "communication" has been considered a standard for effective teaching. Moore, Moore, and McDonald (2008) used a qualitative study to ask questions of 271 college students. The authors found that students cited "learning the material" as the most frequently stated expectation for college courses. Of Marsh's nine dimensions (1982), "Group interaction", "Individual rapport", and "Enthusiasm" could be considered elements of the ability to communicate. Study results are related to findings by Cohen (1981) in his meta-analysis study which indicated instructor skills (r =.50) and class structure (r =.47) were significantly related to overall score.

Facilitation of learning

Facilitation of learning also appeared in all rules and suggests related aspects of communication. Study results agree with Wang et al., (2009) that facilitation of learning is an important factor in the modeling of overall score. The authors found the facilitation item to be important in all six top rules. In their book on facilitative teaching, Wittmer and Myrick (1974) provided instructor characteristics for promotion of learning as: good listeners; empathetic; caring; concerned; genuine; warm; interested; knowledgeable; trusting; friendly with a sense of humor; dynamic; and able to communicate. Carl Rogers (1983) also described a facilitative teacher as one who created a learning environment rather than simply transmitting knowledge. Seldin (1984) believes that being interested in teaching students and motivating students are two of the five characteristics of an effective teacher.

Respect and concern for students

Empathy for students was reflected in all rules used to describe “Excellent” or “Poor” instructors. Caring instructors were rated higher than instructors rated low on the *respect and concern* item. Wang et al. (2009) found *respect and concern* to be a consideration in only one rule out of six when using information from both undergraduate and graduate students. Seldin (1984) refers to “fair and reasonable management” while Chickering and Gamson (1987) use the statement “respects diverse talents and ways of learning.” A supportive climate as defined by consideration and respect was found to be important by Kim et al. (2000).

Instructor’s overall organization of the course

Organization showed up as important in three of the six rules; matching the results of Wang et al. (2009). Rule two reveals that a well organized instructor can overcome lower ratings for respect and concern and still achieve an “Excellent” overall score. Conversely an instructor that is not organized along with not caring about students will generally receive a low overall score. Marsh (1982) and Seldin (1984) both have a measure for organization listed in the important dimensions of effective teaching. Course organization is clearly under the control of the instructor and can be enhanced through professional development.

Instructor’s interest in your learning

Interest in student learning is related to respect and concern and the associated item on the student evaluation form showed up in rules two, five, and six. Wang et al. (2009) found instructor interest in learning to be in one rule for “Excellent” and one rule for “Poor.” Seldin (1984) refers to “fair and reasonable management” and “interested in teaching” as a description of caring about the teaching process. Chickering and Gamson (1987) use the statement “respects diverse talents and ways of learning” which speaks to a supportive course climate.

Instructor’s assessment of your progress

Assessment of progress only showed up in rule three and only after *communication, facilitation of learning,* and *respect and concern for students*. In this rule if an instructor was rated slightly lower on the respect and concern scale but received an “Excellent” or “Very good” for assessment of progress they were more likely to get an overall “Excellent” score. Wang et al. (2009) had this item in two of six important rules. Seldin (1984) refers to fair and reasonable management as a factor for effective teaching. Chickering and Gamson (1987) use “encourages contacts between students and faculty” and “gives prompt feedback” as important principles for undergraduate instruction. Abrami and d’Apollonia (1991) identified “feedback” as an important dimension of effective instruction.

Expression of expectations for performance and stimulation of interest

Expectations of performance and *stimulation of interest* were only present in rule six which describes a “Poor” overall score. These two items are only important after taking into account scores for *communication, facilitation, respect and concern,* and *interest in student learning*. In this rule if an instructor rated “Poor” for both *expectation of student performance* and *stimulation of interest,* in addition to low scores for *communication, facilitation* and *interest in student learning*, they had an elevated chance of getting an overall “Poor” score. Wang et al. (2009) did not identify these items as important when using undergraduate and graduate student responses. This difference between this study and Wang et al. (2009) may be related to undergraduates and graduate students having more college experience and therefore having the instructor spell out what is necessary for performance in the course is not critical to their rating of the instructor or they are able to recognize when instructors are presenting performance expectations. Seldin (1984) refers to fair and reasonable management and motivating students as two important components of effective teaching. Marsh (1982) refers to enthusiasm as a dimension used to evaluate effective teaching. Abrami and d'Apollonia (1991) designate "rapport" as a dimension of effective instruction. Remedios and Lieberman (2005) found student perceived quality of teaching was related to how much students enjoyed or felt stimulated by the course content.

Limitations

Student information was collected prior to the study and not specifically for the purposes of this study. Procedural conditions were not controlled and direct contact with participants was not possible. The data set available was not a random sample of students or courses. Information was collected from students willing to complete the questionnaire; this group of students may represent a biased sample. Under the current data collection procedures, it was not possible to obtain random samples or complete information from all students enrolled in the course.

There may be important components of the variation in evaluation scores that were not being considered as part of this study. Individual student information was not obtained at the time of data collection because the forms are anonymous. Faculty information could be obtained, but for reasons of confidentiality, instructor information was not used in this study.

Information on how students interpret the items on the student evaluation form and what specific instructor actions were being applied to that item were not available at the present time. Thus, an item such as organization of the course is based on the student perspective of that particular item.

Results could be generalized to other student populations at similar universities or to future semesters at the same university. Results should be most closely related to future semesters at the same institution, as the students, instructors, and courses will continue to be very closely related to the information used in the analysis. The novel methodology developed here however should transfer to other institutions with similar student evaluation information.

Summary

GEP courses taught from September 2002 through January 2009 had an average response rate of 60.34% after removing responses that did not indicate an overall instructor score. Based on this data set, 43.82% of students regarded their instructor as “Excellent” with only 2.26% indicating their overall rating belonged in the “Poor” category.

Summary results indicated that the overall evaluation score for instructors of GEP courses appeared to benefit from having a common theme. Percentage of “Excellent” scores increased from 42.24% to 45.78% for semesters with the common theme implemented. Face-to-face courses had a higher percentage of “Excellent” scores (44.12%) as compared to all other instructional modes (41.30%). Foundational areas differed considerably in the percent of “Excellent” scores given by students. The results showed that mathematics foundational area courses had the lowest proportion of “Excellent” overall scores (35.27%) of any foundational area. Communications (46.42%) and historical and cultural (46.40%) foundation areas had the highest proportion of “Excellent” overall scores.

Data mining methodology, specifically decision tree analysis was incorporated to understand the relationship of overall instructor scores with other items and course factors. Decision trees are very efficient modeling tools in situations with large data sets where there are multiple variables with missing values. Missing data values do not need to be excluded or imputed in order for modeling procedures to find important variables that will differentiate values of the overall evaluation score. Results are in the form of easily related rules for interpreting values of the overall evaluation score. The rules can be assessed using accuracy information in the form of odds ratios and misclassification rates.

Research findings reflect that students had a consistent approach to completion of the evaluation forms and specific individual items on the form were related to important components in the students' perception of instruction. Items related to communication, facilitation, organization, respect and concern, instructor's assessment of your progress in the course, instructor's interest in your learning and stimulation of interest in the course were found to characterize aspects important to students in regard to evaluating instructors. Course variables of class size, foundational area, GEP theme, and percent responding did not change student approach to evaluation.

Effective instruction has numerous dimensions that comprise factors important for general teaching environments. These principles may have different terminology depending on whether Marsh (1982) or Chickering and Gamson (1987) or Abrami and d'Apollonia (1991) are defining the dimensions. Kolitch and Dean (1999) state the goal of effective teaching to be:

> "The aim of the 'effective' teacher is to transfer the subject matter to the students through a clear and organized presentation of ideas, raising interesting questions, and using relevant examples. It is the instructor's responsibility to capture students' attention and to foster their interest in the subject matter by being stimulating, eloquent, and dynamic." (p. 33).

The study results showed a striking resemblance to important aspects highlighted by Kolitch and Dean (1999). Transferring information is synonymous with *communication of ideas and information* and *facilitation of learning*, which appeared in all data mining rules identifying "Excellent" or "Poor" instructors. Additionally, *instructor's overall organization of the course*, *respect and concern for students*, and *stimulating interest* showed up as important items for determining the overall score. *Assessment of progress* and *expectations of performance* are related to communication of expectations or classroom management skills.

Study results generally agree with Wang et al. (2009) on the important items related to modeling overall evaluation scores. *Communication, facilitation, organization,* and *respect and concern* were consistent items in data mining rules in both studies. Students evaluating instructors of GEP courses rated items regarding *expectation of performance, assessment of their progress, interest in student learning,* and *stimulation of interest* as additional important components of an effective instructor.

Implications

These findings suggest that students reward instructors who they perceive as organized and strive to clearly communicate course content. These characteristics can be improved through mentoring or professional development workshops for instructors. Additionally, instructors need to be informed that students connect respect and concern and having an interest in student learning with the overall score they give the instructor. Finally, assessment of student progress and expectation of performance are important components of classroom management that can be easily improved by instructors through their own efforts or with the assistance of colleagues. Understanding how students perceive instruction will benefit the learning environment for all stakeholders.

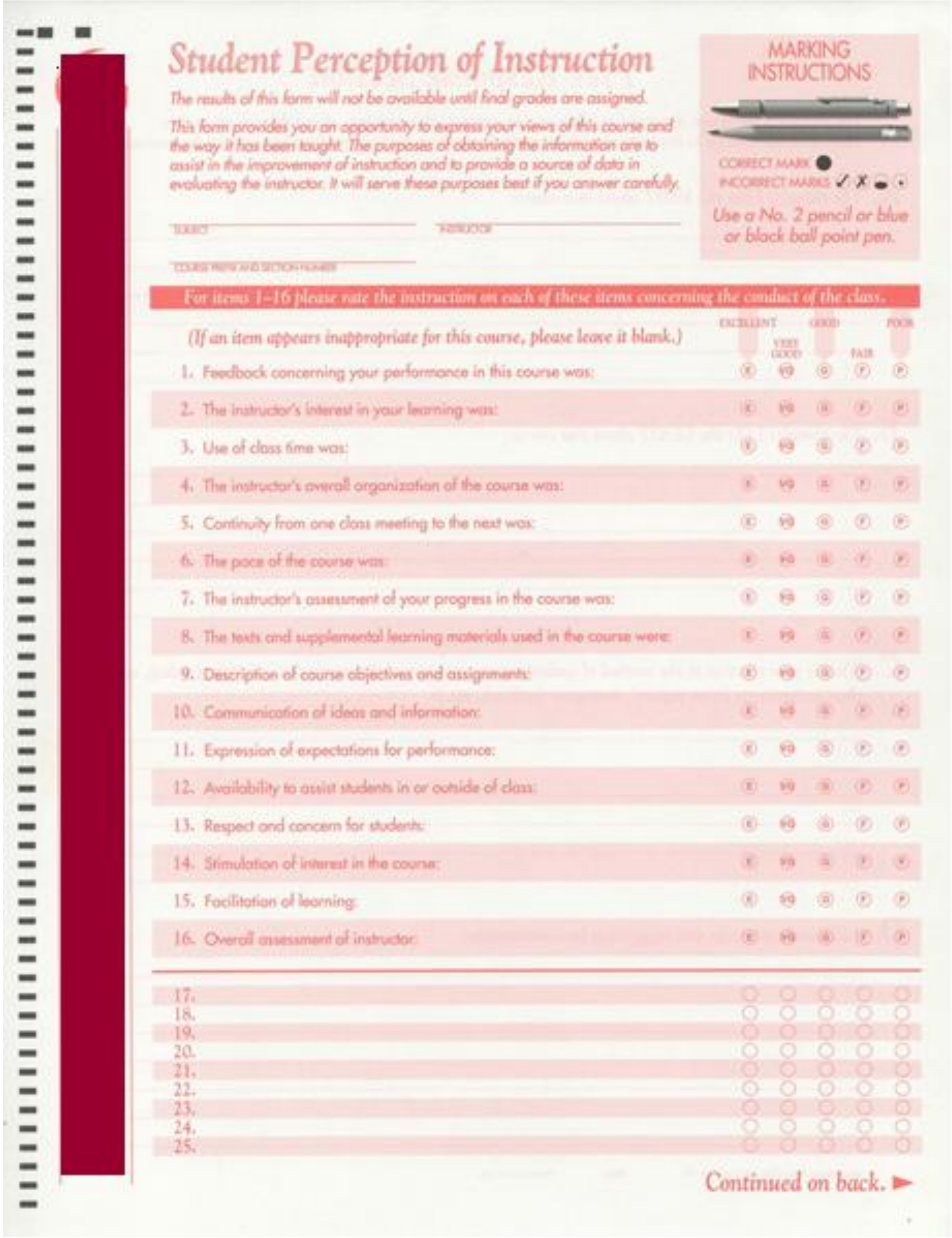
The characteristics of this study are only a small proportion of potentially important factors in higher education learning environments. The lack of grade, student demographic, and instructor demographic information eliminates aspects of the learning environment that may be related to the student evaluation scores. Follow up studies should be done to assess additional variables not included in the present study. Critical issues of student learning (e.g., test scores, assessment measures) were also not considered as part of the present study.

Recommendations for Future Studies

Although research on student evaluation data is quite extensive, there are areas of study that are incomplete or very limited. Recent analysis tools such as decision trees can enhance the understanding of student evaluation data. The following areas of potential research are listed:

1. Conduct this study using additional information such as grade distributions.
2. Modify data set construction to use a matched-pair design, similar to medical studies, so that confounding factors can be isolated.
3. Compare results of this study to all courses using the same time period in order to identify differences between all courses (graduate and undergraduate) with GEP courses.
4. Combine data sets from previous work done by Wang et al. with more recent evaluation information to determine whether scores are changing over time.
5. Expand Cohen’s 1981 meta-analysis study to include recent results.

APPENDIX A
STUDENT EVALUATION FORM



[redacted]

Student Perception of Instruction

The results of this form will not be available until final grades are assigned.

This form provides you an opportunity to express your views of this course and the way it has been taught. The purposes of obtaining the information are to assist in the improvement of instruction and to provide a source of data in evaluating the instructor. It will serve these purposes best if you answer carefully.

MARKING INSTRUCTIONS

CORRECT MARK ●
INCORRECT MARKS ✓ ✗ ◒ ⊙

Use a No. 2 pencil or blue or black ball point pen.

SUBJECT ______________ INSTRUCTOR ______________

COURSE PREFIX AND SECTION NUMBER ______________

For items 1–16 please rate the instruction on each of these items concerning the conduct of the class.

(If an item appears inappropriate for this course, please leave it blank.)

	EXCELLENT	VERY GOOD	GOOD	FAIR	POOR
1. Feedback concerning your performance in this course was:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
2. The instructor's interest in your learning was:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
3. Use of class time was:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
4. The instructor's overall organization of the course was:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
5. Continuity from one class meeting to the next was:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
6. The pace of the course was:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
7. The instructor's assessment of your progress in the course was:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
8. The texts and supplemental learning materials used in the course were:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
9. Description of course objectives and assignments:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
10. Communication of ideas and information:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
11. Expression of expectations for performance:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
12. Availability to assist students in or outside of class:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
13. Respect and concern for students:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
14. Stimulation of interest in the course:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
15. Facilitation of learning:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ
16. Overall assessment of instructor:	Ⓔ	VG	Ⓖ	Ⓕ	Ⓟ

17.	○	○	○	○	○
18.	○	○	○	○	○
19.	○	○	○	○	○
20.	○	○	○	○	○
21.	○	○	○	○	○
22.	○	○	○	○	○
23.	○	○	○	○	○
24.	○	○	○	○	○
25.	○	○	○	○	○

Continued on back. ►

APPENDIX B
GENERAL EDUCATION COURSE LIST

GEP Summary				Year						
Communication Foundations									Future	
Class	01-02	02-03	03-04	04-05	05-06	06-07	07-08	08-09	09-10	
ENC1101	x	x	x	x	x	x	x	x	x	English Composition I
ENC1102	x	x	x	x	x	x	x	x	x	English Composition II
SPC1600C	x	x	x	x	x	x	x	x	SPC1608	Fund. of Oral Communication
SPC1016	x	x	x	x	x	x	x	x	SPC1603	Fund. of Technical Presentation
COM1000					x	x	x	x	x	Communications
Cultural and Historical Foundations										
EUH2000/2001	x	x	x	x	x	x	x	x	x	Western Civilization I & II
HUM2211/2230	x	x	x	x	x	HUM2210	x	x	x	Humanistic Tradition I & II
AMH2010/2020	x	x	x	x	x	x	x	x	x	U.S. History 1492-1877&1877-present
WOH2012/2022	x	x	x	x	x	x	x	x	x	World Civilization I & II
ARH2050	x	x	x	x	x	x	x	x	x	The History of Art I
ARH2051	x	x	x	x	x	x	x	x	x	The History of Art II
MUL2010	x	x	x	x	x	x	x	x	x	Enjoyment of Music
THE1020	x	THE2000	x	x	x	x	x	x	x	Theatre Survey
FIL1001	x	x	x	x	x	x	x	FIL1000	x	Cinema Survey
REL2300	x	x	x	x	x	x	x	x	x	World Religions
PHI2010	x	x	x	x	x	x	x	x	x	Introduction of Philosophy
LIT2110	x	x	x	x	x	x	x	x	x	World Literature I
LIT2120	x	x	x	x	x	x	x	x	x	World Literature II

GEP Summary (Continued)

						Year				
Communication Foundations									Future	
Class	01-02	02-03	03-04	04-05	05-06	06-07	07-08	08-09	09-10	
MGF1106	x	x	x	x	x	x	x	x	x	Finite Mathematics
CGS1060C	x	x	x	x	x	x	x	x	x	Introduction to Computer Science
STA1060C	x	x	x	x	x	x	x	x	x	Basic Statistics using Microsoft Excel
STA2014C	x	x	x	x	x	x	x	x	x	Principles of Statistics
Social Foundations										
ECO2013	x	x	x	x	x	x	x	x	x	Principles of Economics I
ECO2023	x	x	x	x	x	x	x	x	x	Principles of Economics II
POS2041	x	x	x	x	x	x	x	x	x	American National Government
PSY2013	x	PSY2012	x	x	x	x	x	x	x	General Psychology
SYG2000	x	x	x	x	x	x	x	x	x	General Sociology
ANT2000	x	x	x	x	x	x	x	x	x	General Anthropology
Science Foundations										
AST2002	x	x	x	x	x	x	x	x	x	Astronomy
PSC1121	x	x	x	x	x	x	x	x	x	Physical Science
PHY2053C	x	x	x	x	x	x	x	x	x	College Physics
CHM1020	x	x	x	x	x	x	x	x	x	Concepts in Chemistry
BSC1005	x	x	x	x	x	x	x	x	x	Biological Principles
BSC1050	x	x	x	x	x	x	x	x	x	Biology and Environment
GLY1030	x	x	x	x	x	x	x	x	x	Geology & Its Applications
GEO1200	x	x	x	x	x	x	x	x	x	Physical Geography
ANT2511	x	x	x	x	x	x	x	x	x	The Human Species
MCB1310				x	x	x*	x	x	x	Biotechnology and Genetics

APPENDIX C
SPEARMAN CORRELATION MATRIX

*Item Correlation Matrix for the Student Perception of Instruction Form**

	Item number														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	.70														
3	.57	.66													
4	.61	.65	.75												
5	.57	.63	.71	.76											
6	.60	.63	.65	.67	.69										
7	.77	.71	.61	.65	.63	.68									
8	.51	.53	.54	.55	.54	.56	.58								
9	.63	.66	.64	.70	.67	.67	.69	.62							
10	.63	.71	.70	.73	.69	.69	.69	.58	.76						
11	.68	.71	.64	.68	.66	.67	.74	.57	.74	.77					
12	.62	.67	.56	.60	.59	.58	.67	.51	.63	.64	.68				
13	.61	.73	.60	.62	.61	.61	.67	.50	.65	.69	.70	.72			
14	.60	.70	.66	.65	.64	.66	.66	.56	.66	.74	.69	.62	.69		
15	.64	.71	.68	.70	.68	.69	.70	.59	.70	.76	.74	.66	.71	.78	

Number of observations differs for each comparison due to missing data values.

Minimum number of responses is 287,868.

REFERENCES

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In P. Abrami, M. Theall, & L. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them? New Directions for Institutional Research, no. 109,* (pp. 59-87). San Francisco: Jossey-Bass.

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness--generalizability of" N= 1" research: Comment on Marsh (1991). *Journal of Educational Psychology, 83*(3), 411-415.

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*(2), 219-231.

Abrami, P. C., Theall, M., & Mets, L. M. (Eds.). (2001). The student ratings debate: Are they valid? How can we best use them? *New Directions for Institutional Research, no. 109,* (pp. 1-6). San Francisco: Jossey-Bass.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*(2), 153-166.

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science, 9*(1), 67-84.

Algozzine, B., Gretes, J., Flowers, C., Howley, L., Beattie, J., Spooner, F., et al. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching, 52*(4), 134-141.

Aloi, S. L., Gardner, W. S., & Lusher, A. L. (2003). A framework for assessing general education outcomes within the majors. *Journal of General Education, 52*(4), 237–252.

Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni-and multidimensional models. *Studies in Higher Education, 30*(6), 723-748.

Association of American Colleges and Universities. (2008). *Project on accreditation and assessment.* Retrieved from the Association of American Colleges and Universities Web site: http://www.aacu.org/resources/liberaleducation/accreditation.cfm

Awbrey, S. M. (2005). General education reform as organizational change. *The Journal of General Education, 54*, 1-21.

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education, 48*(3), 193-210.

Benton, T. H. (2008). Do students' online ratings of courses 'suck' (or 'rock')? *Chronicle of Higher Education, 55*(11), A49-A52. Retrieved from http://ezproxy.lib.ucf.edu/login?URL=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=35536396&site=ehost-live

Beyers, C. (2008). The hermeneutics of student evaluations. *College Teaching, 56*(2), 102-106.

Brandenburg, D. C., Slinde, J. A., & Batista, E. E. (1977). Student ratings of instruction: Validity and normative interpretations. *Research in Higher Education, 7*(1), 67-78.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1998). *Classification and regression trees*. Boca Raton, Fla.: Chapman & Hall/CRC.

Campbell, J. P., & Bozeman, W. C. (2008). The value of student ratings: Perceptions of students, teachers, and administrators. *Community College Journal of Research and Practice, 32*(1), 13-24.

Campbell, J. P. (2005). Evaluating teacher performance in higher education: The value of student ratings. *Dissertation Abstracts International*, 66(08), 2851A.(UMI No. AAT 3188108).

Carle, A. C. (2009). Evaluating college students' evaluations of a professor's teaching effectiveness across time and instruction mode (online vs. face-to-face) using a multilevel growth modeling approach. *Computers & Education, 53,* 429-435.

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. *In P. Selden (Ed.), Changing Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and promotion/tenure Decisions. Bolton, MA: Anker.*

Cashin, W. E., & Clegg, V. L. (1987). *Are students ratings of different academic fields different?* Paper presented at the *Annual Meeting of the American Educational Research Association, New Orleans, LA.* (ERIC Document Reproduction Service no. ED 286 935).

Cashin, W. E., & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology, 84*, 563-563.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, *44*(5), 495-518.

Centra, J. A. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education, 71*(1), 17-33.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness.* San Francisco: Jossey-Bass.

Chang, T. S. (2000). *An application of regression models with student ratings in determining course effectiveness.* Paper presented at the *Annual Meeting of the American Educational Research Association, New Orleans, LA.*(ERIC Document Reproduction Service no. ED 455 311).

Chickering, A. W., & Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *AAHE Bulletin, 39*(7), 3-7.

Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education, 28*(2), 149-160.

Cohen, P. A. (1982). Validity of student ratings in psychology courses: A research synthesis. *Teaching of Psychology, 9*(2), 78-82.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*(3), 281.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, *41*(5), 511-535.

Council for Higher Education Accreditation. (2008). *An overview of U.S. accreditation, May 2009.* Retrieved from the Council for Higher Education Accreditation Web site http://www.chea.org/

Crumbley, L., Henry, B. K., & Kratchman, S. H. (2001). Students' perceptions of the evaluation of college teaching. *Quality Assurance in Education*, *9*(4), 197-207.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, *52*(11), 1198-1208.

Dillman, D. A. (2000). *Mail and electronic surveys: The tailored design method* (2nd ed.). New York: John Wiley & Sons.

Dooris, M. J. (1997). *Instructor personality and the politics of the classroom.* Retrieved 2/25, 2010, from http://www.mnsu.edu/psych/Damron_politics.html

Dziuban, C. D., Wang, M. C., & Cook, I. J. (2004). *Dr. fox rocks: Student perceptions of excellent and poor college teaching.* Unpublished manuscript.

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education*, *43*(4), 483-501.

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II—Evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34*(2), 151-211.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*(6), 583-645.

Feldman, K. A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education, 24*(2), 139-213.

Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education, 10*(2), 149-172.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education, 9*(3), 199-242.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education, 5*(3), 243-288.

Franklin, J., & Theall, M. (1989). *Who reads ratings: Knowledge, attitude, and practice of users of student ratings of instruction.* Paper presented at the *Annual Meeting of the American Educational Research Association, San Francisco, CA*. (ERIC Document Reproduction Service no. ED 306 241).

Glynn, S. M., Aultman, L. P., & Owens, A. M. (2005). Motivation to learn in general education programs. *Journal of General Education, 54*(2), 150–170.

Goyder, J. (2009). *Report on UW faculty of arts experiment with course evaluation administered by web.* Retrieved from: http://aco.uwaterloo.ca/report_on_experiment.pdf

Gravestock, P., & Gregor-Greenleaf, E. (2008). Student course evaluations: Research, models and trends. *Higher Education Quality Council of Ontario.* Ontario: Queens Printer.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182-1186.

Greenwald, A. G., & Gillmore, G. M. (1997). No pain, no gain? the importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89,* 743-751.

Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology, 29*(4), 410-425.

Grimes, P. W., Millea, M. J., & Woodruff, T. W. (2004). Grades—Who's to blame? student evaluation of teaching and locus of control. *The Journal of Economic Education, 35*(2), 129-147.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, Mass.: MIT Press.

Hativa, N., & Birenbaum, M. (2000). Who Prefers What? disciplinary differences in students' preferred approaches to teaching and learning styles. *Research in Higher Education, 41*(2), 209-236.

Heckert, T. M., Latier, A., Ringwald-Burton, A., & Drazen, C. (2006). Relations among student effort, perceived class difficulty appropriateness, and student evaluations of teaching: Is it possible to" buy" better evaluations through lenient grading? *College Student Journal, 40*(3), 588-596.

Joe, J. N., Harmes, J. C., & Barry, C. L. (2008). Arts and humanities general education assessment: A qualitative approach to developing program objectives. *The Journal of General Education, 57*(3), 131-151.

Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education, 5*(4), 419-434.

Kendall, M. G., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed.). London: Charles Griffin.

Kim, C., Damewood, E., & Hodge, N. (2000). Professor attitude: Its effect on teaching evaluations. *Journal of Management Education, 24*(4), 458-473.

Koh, H. C., & Tan, T. M. (1997). Empirical investigation of the factors affecting SET results. *International Journal of Educational Management, 11*(4), 170-178.

Kolitch, E., & Dean, A. (1999). Student ratings of instruction in the USA: Hidden assumptions and missing conceptions aboutgoods teaching. *Studies in Higher Education, 24*(1), 27-42.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. In P. Abrami, M. Theall, & L. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them? New Directions for Institutional Research, no. 109,* (pp. 9-25). San Francisco: Jossey-Bass.

Landrum, R. E., & Braitman, K. A. (2008). The effect of decreasing response options on students' evaluation of instruction. *College Teaching, 56*(4), 215-218.

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99*(4), 775-790.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*(5), 707-754.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75*(1), 150-166.

Marsh, H. W. (1982). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology, 74*(2), 264-279.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979b). Class size, students' evaluations, and instructional effectiveness. *American Educational Research Journal, 16*(1), 57-70.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979a). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. *Journal of Educational Psychology, 71*(2), 149-160.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187-1197.

McClave, J. T., & Sincich, T. (2006). *Statistics* (Eleventh ed.) Upper Saddle River NJ: Prentice Hall.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*, 1218-1225.

McKeachie, W. J. (1987). Instructional evaluation: Current issues and possible improvements. *The Journal of Higher Education, 58*(3), 344-350.

Moore, M. L., Moore, R. S., & McDonald, R. (2008). Student characteristics and expectations of university classes: A free elicitation approach. *College Student Journal, 42*(1), 82-89.

Moore, S., & Kuol, N. (2005). Students evaluating teachers: Exploring the importance of faculty reaction to feedback on teaching. *Teaching in Higher Education, 10*(1), 57-73.

National Center for Educational Statistics, U.S. Department of Education. (2009). Federal support and estimated federal tax expenditures for education, by category. In *Digest of education statistics 2009* (table 373). Retrieved October 21, 2009, from the National Center for Education Statistics Web site:

http://nces.ed.gov/programs/digest/d08/tables/dt08_373.asp/

Ornstein, A. C. (1995). Beyond effective teaching. *Peabody Journal of Education, 70*(2), 2-23.

Ory, J. C., Braskamp, L. A., & Pieper, D. M. (1980). Congruency of student evaluative information collected by three methods. *Journal of Educational Psychology, 72*(2), 181-185.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In P. Abrami, M. Theall, & L. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them? New Directions for Institutional Research, no. 109,* (pp. 27-44). San Francisco: Jossey-Bass.

Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology, 72*(3), 321-325.

Paulsen, M. B. (2002). Evaluating teaching performance. *New Directions for Institutional Research, no. 114,* (pp. 5-18). San Francisco: Jossey-Bass.

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile? *Quality Assurance in Education, 15*(2), 178-191.

Radmacher, S. A., & Martin, D. J. (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *The Journal of Psychology, 135*(3), 259-268.

Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal, 34*(1), 91-115.

Roche, L. A., & Marsh, H. W. (2000). Multiple dimensions of university teacher self-concept. *Instructional Science, 28*(5), 439-468.

Rogers, C. (1983). Freedom to learn for the eighties. *Columbus OH, Charles Merrin,*

Saroyan, A., & Amundsen, C. (2001). Evaluating university teaching: Time to take stock. *Assessment &# 38; Evaluation in Higher Education, 26*(4), 341-353.

SAS Institute. (2009). SAS © Enterprise Miner™ Version 6.1 [computer software]. Cary, NC: Author.

SAS Institute. (2009). SAS Users Guide, version 8 [available via software]. Cary, NC: Author.

Scriven, M. (1983). Evaluation ideologies. *Evaluation Models,* 249-278.

Seldin, P. (1999). *Changing Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and promotion/tenure Decisions.* Bolton, MA: Anker.

Seldin, P. (1984). Faculty evaluation: Surveying policy and practices. *Change,* 28-33.

Shao, L., Anderson, L., & Newsome, M. (2007). Evaluating teaching effectiveness: Where we are and where we should be. *Assessment &# 38; Evaluation in Higher Education, 32*(3), 355-371.

Snyder, T. D., Dillow, S., & Hoffman, C. M. (2009). Digest of education statistics 2008.

Snyder, T. D., & Dillow, S. A. (2010). *Digest of education statistics 2009*

Sprinkle, J. E. (2008). Student perceptions of effectiveness: An examination of the influence of student biases. *College Student Journal, 42*(2), 276-293.

Stark-Wroblewski, K., Ahlering, R. F., & Brill, F. M. (2007). Toward a more comprehensive approach to evaluating teaching effectiveness: Supplementing student evaluations of teaching with pre-post learning measures. *Assessment & Evaluation in Higher Education, 32*(4), 403-415.

Sudermann, D. P. (1992). Toward a definition of core curriculum. Retrieved from (ERIC Document Reproduction Service no. ED 351 951).

Summers, J. J., Waigandt, A., & Whittaker, T. A. (2005). A comparison of student achievement and satisfaction in an online versus a traditional face-to-face statistics class. *Innovative Higher Education, 29*(3), 233-250.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? In P. Abrami, M. Theall, & L. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them? New Directions for Institutional Research, 109,* 45-56.

Thomas, E. H., & Galambos, N. (2004). What satisfies students? mining student-opinion data with regression and decision tree analysis. *Research in Higher Education, 45*(3), 251-269.

Thompson Jr, R. J., & Serra, M. (2005). Use of course evaluations to assess the contributions of curricular and pedagogical initiatives to undergraduate general education learning objectives. *Education, 125*(4), 693-702.

Toby, S. (1993). Class size and teaching evaluation: Or, the "general chemistry effect" revisited. *Journal of Chemical Education, 70*(6), 465-466.

Trout, P. (2000). Flunking the test: The dismal record of student evaluations. *Academe, 86*(4), 58-61.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, *23*(2), 191-211.

Walumbwa, F. O., Wu, C., & Ojode, L. A. (2004). Gender and instructional outcomes. *Journal of Management Development, 23*(2), 124-140.

Wang, M. C. (2007). Data mining I course notes.

Wang, M. C., Dziuban, C. D., Cook, I. J., & Moskal, P. D. Dr. fox rocks: Using data-mining techniques to examine student ratings of instruction. *Quality Research in Literacy and Science Education*, 383-398.

Wittmer, J., & Myrick, R. D. (1974). *Facilitative teaching: Theory and practice*. Pacific Palisades, CA: Goodyear Publishing.