

NOTE TO USERS

This reproduction is the best copy available.

UMI[®]

BOSTON UNIVERSITY
SCHOOL OF EDUCATION

Dissertation

COMPARING INSTRUCTOR SELF-PERCEPTION VERSUS STUDENT
PERCEPTIONS USING THE SAME TEACHING EVALUATION INSTRUMENT:
A STUDY OF COMPUTER SCIENCE COURSES
IN AN URBAN MASTER'S DEGREE PROGRAM

by

LAURIE SCHWARTZ NAPARSTEK

B.S. University of Massachusetts/Amherst, 1981
M.S. Suffolk University, 1988
Ed.M. Harvard University, 1989

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Education

2005

UMI Number: 3157400

Copyright 2005 by
Naparstek, Laurie Schwartz

All rights reserved.

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI[®]

UMI Microform 3157400

Copyright 2005 by ProQuest Information and Learning Company.

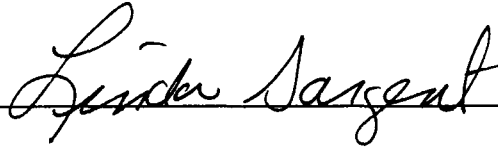
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

© Copyright by
LAURIE SCHWARTZ NAPARSTEK
2005

Approved by

First Reader



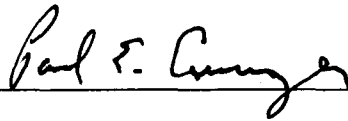
Linda Sargent, Ed.D.
Assistant Clinical Professor (1990 – 2003)

Second Reader



Leonard Zaichkowsky, Ph.D.
Professor of Education

Third Reader



Paul Cournoyer, Ph.D.
Assistant Professor of Computer Science (1993 –2000)

Fourth Reader



Christopher Sotak, Ph.D.
Professor and Department Head of Biomedical Engineering,
Worcester Polytechnic Institute

DEDICATION

To my husband Jay, my mother and father, my brother Eric, my nephew Ben and
in memory of my brother Ken.

ACKNOWLEDGMENTS

My appreciation to my husband, family and friends for their ongoing support.

I was most fortunate to have a dedicated, patient and inspirational committee including Dr. Linda Sargent and Dr. Leonard Zaichkowsky. To my third reader, Dr. Paul Cournoyer, thank you for your unwavering encouragement. My deepest appreciation goes to my fourth reader, Dr. Christopher Sotak, Professor and Department Head of Biomedical Engineering at Worcester Polytechnic Institute, without whose time and expertise this project would not have been completed. Dr. Sotak assisted significantly with the interpretation of the data results and the organization of this dissertation.

My gratitude also to Romer Rosales who was a graduate student in computer science at the time of the study. He meticulously created and input the raw data into the spreadsheets and ran the preliminary data analyses. Also, my appreciation to Dr. Janice Weinberg, Sc.D. in Biostatistics and Assistant Professor of Biostatistics, Mathematics and Statistics at Boston University, who, based on consultations with the author, advised and conducted the statistical analyses on the data. Also, my thanks to Oge Harrison, Niyi Taiwo, Judith Nordberg, Paula Gopin and Kelly Pace.

**COMPARING INSTRUCTOR SELF-PERCEPTION VERSUS STUDENT
PERCEPTIONS USING THE SAME TEACHING EVALUATION INSTRUMENT:
A STUDY OF COMPUTER SCIENCE COURSES
IN AN URBAN MASTER'S DEGREE PROGRAM**

(Order No.)

LAURIE SCHWARTZ NAPARSTEK

Boston University School of Education, 2005

Major Professor: Linda Sargent, Ed.D. Assistant Clinical Professor (1990-2003)

ABSTRACT

This study compares instructor self-perceptions with student perceptions of teaching quality using the same 16-item evaluation instrument. Three hypotheses were investigated: (1) Instructors' self-evaluations will be higher than those of their respective students; (2) The more similar student-instructor perceptions, the more likely instructors will receive a higher score compared to when student-instructor perceptions are more divergent; and (3) Students taking a course as a major requirement will be more critical of the instructor than those students taking the course as a distribution requirement or an elective.

A total of 1,524 individuals (1,452 graduate students and 72 instructors) in a part-time evening computer science program participated in the study of 79 courses over the spring and fall semesters of 1996. Overall, instructors generally perceived themselves more positively than their students, although statistically significant differences were observed for only three relevant items (involving

grading fairness, presentation clarity and instructor enthusiasm) of the 16 items evaluated. Instructors whose perceptions were more similar to their students were generally rated higher than those instructors whose perceptions were more divergent from their students; however, the difference was not significant. Finally, contrary to the third hypothesis, the reason for taking a course did not have a significant effect on student ratings of the instructor.

Table of Contents

Chapter I: Introduction

A. Background of the Problem	1
B. Rationale and Significance of the Study	4
C. Statement of the Problem	15
D. Research Questions	23
E. Hypotheses	23
F. Definition of Terminology	25
G. Summary and Overview	26

Chapter II: Review of the Literature

A. Introduction	29
B. Historical Background of Student Evaluations	29
C. Reliability and Validity of Evaluations	35
D. Design, Protocols and Usages of Evaluations	42
E. Comparative Studies of Student and Instructor Evaluations	50
F. Summary of the Relevant Research	59

Chapter III: Methodology

A. Introduction	61
B. Sample Population to be Studied	61
C. Instrumentation	62
D. Data Collection Procedures	63
E. Research Design	65

F. Treatment of Data	66
G. Limitations of the Methodology	71
Chapter IV: The Results	
A. Introduction	73
B. Descriptive Statistics/Data Analysis related to Hypothesis One	73
C. Descriptive Statistics/Data Analysis related to Hypothesis Two	83
D. Descriptive Statistics/Data Analysis related to Hypothesis Three	85
Chapter V: Discussion	
A. Overview of the Study	88
B. Overview of the Results Related to the Research Hypotheses	90
C. Summary of the Research Results	114
D. Limitations of the Study	115
E. Implications of the Research - A Multi-Medium Approach	123
F. Suggestions for Future Research	129
G. Summary	135
Appendices	
Appendix A: Course Evaluation Form	141
Appendix B: Informed Consent Form	142
Appendix C: Approval Memo from College Dean	143
Graphs	144
References	185
Curriculum Vitae	192

List of Tables

Table 1: Student and Instructor Mean Scores (± 1 S.D), Based on a 5-Point Likert Scale, for the 16 Items on the Teaching Evaluations	74
Table 2: Instructor and Student Mean Scores and Mean, Median, Standard Deviation (S.D), Range and <i>P</i> value of the Differences between Instructor and Average Student Likert Scores for the 16 Items on the Teaching Evaluations	77
Table 3: Results from Clustered Data Analysis Comparing Instructor and Student Likert Scores for the 16 Items on the Teaching Evaluation	80
Table 4: Discrepancy Analysis of Difference between Instructor and Student Likert Scores for the 16 Items on the Teaching Evaluations	82
Table 5: Mean (± 1 S.D.) Instructor and Student Likert Scores for Statistically Similar and Dissimilar Groups of Items in Table 2	85
Table 6: Descriptive Statistics on the Difference between Teacher and Student Rating Based on Reason for Taking the Course	86

List of Figures

Figure 1: Instructor and Student Mean Scores (± 1 S.D.), Based on a 5-Point Likert Scale, for the 16 Items on the Teaching Evaluations	75
--	----

CHAPTER I

INTRODUCTION

Background of the Problem

The following dissertation is a study of a comparative nature examining the relationship between students and their respective teachers according to course evaluations scores in a graduate school program. The act of teaching is an important concept to be defined at the beginning of this study. Anderson and Burns (1989) present a definition of the word “teacher” as it appears in the Dictionary of Education (Good, 1973). They define the role of teachers as “those persons who are employed in schools in an official capacity for the expressed purpose of leading the learning process” (p. 3). Another definition states that teaching “comes from the integrity of the teacher, from his or her relation to subject and students, from the capricious chemistry of it all” (Palmer, 1990, p. 11).

The characteristics that comprise the model instructor are another concept worthy of attention early in this paper. Dulz and Lyons (2000) write, “a picture of the ‘ideal’ instructor emerges [as a] ‘sparkplug’, a motivator, a source of knowledge, the center, the linchpin of the learning process...also known as the ‘sage on the stage’” (What Gets Measured section, ¶4). Dulz and Lyons continue to write that an instructor is “assumed to be the expert and the primary source of knowledge” (Conclusion section, ¶3).

What makes an instructor effective? What measures exist to determine if a teacher adequately conveys his/her knowledge to the classroom in which they teach? What attributes should instructors possess in order to positively relate to their students? These are questions that have been of interest to this author during much of her academic life.

Kaufman (1981) outlines the primary tool of measurement with which to answer the aforementioned questions. Kaufman writes,

“among the sources of information on teacher effectiveness are systematic ratings made by students [that are] usually paper and pencil measures called teacher evaluation scales, and consist of items related to specific characteristics which are thought to be essential to effective teaching” (p. 2)

One way that teachers' responsibilities are reviewed is through the student evaluation process. Evaluations (generally completed by the students) “identify those characteristics or qualities that set excellent or effective teachers apart from other teachers” (Anderson & Burns, 1989, p. 5). Student evaluations vary in their structure and content. These variables can include a variety of items ranging from knowledge of the subject being taught to the level of instructor enthusiasm to being on time to the classroom (Anderson & Burns).

The author of this dissertation has a Bachelor's degree and two Master's degrees in the fields of Education and Human Development. The author has had a lifelong interest in the processes involved in communication and learning that has culminated in the pursuance of a Doctoral degree and in this study, which combines the two areas of interest exploring the student-teacher relationship. In

general, the author has been interested in the qualities that constitute a positive communicative experience in any relationship, but most specifically in relationships in academic settings. Employed at a large urban university, the author was assigned to the position of liaison supporting approximately seven full-time faculty members and 60 adjunct instructors in an evening, part-time, Computer Science master's degree program. It was during this six-year professional experience that it became abundantly clear that a flaw existed in one measurement of the faculty-student relationship—the course evaluation system.

While overseeing, from an administrative perspective, the instructor evaluation program for the Computer Science Master's Degree program, the author became aware of the powerful impact of the evaluation results both on the professional and personal levels for each instructor. Should the result be reviewed "positively," then the instructor was secured for future classroom assignments. Should the result turn out to be "negative," the author was required to schedule a meeting with the Department Chairman for a consultation session and possible termination. There was also the relief related to positive scores and the inevitable disappointment if the scores were less than satisfactory. It also occurred to the author that the approach was "one-sided" to the exclusion of any party other than the student for comparative feedback.

Over time, it became clear that the intervention, or variable, that could be altered in this particular situation was to involve the instructors themselves in the

evaluative process. The idea emerged that instructors could complete the identical evaluation form used by their classroom students and comparative data could be examined to yield new, and hopefully, enlightening data. With the population readily available, and ultimately eager and willing to participate in exploring the evaluative process within the department, the study was begun in January 1996 over two consecutive semesters. The data were subsequently collected, correlated and examined and will be presented in detail in later chapters of this paper.

Rationale and Significance of the Study

One can argue that student evaluations may or may not be useful for assessing teaching quality. The Center for Teaching and Learning (1994, p. 1) writes, "student evaluations are the most commonly used method of assessing an instructor's effectiveness in the classroom." Hiltner and Loyland (1998) write, "universities are facing pressures to provide evidence on how effectively they [the institutions] are accomplishing their missions" (Abstract section). Instructors and their respective academic environments are held accountable in their teaching practices. White (1995) writes that many academic departments use student evaluation instruments even though they "believed that [they were], at best, an imperfect tool for measuring effectiveness" (p. 84). White continues, "departments are devoting increased time and resources to this issue-despite their discomfort with the assessment process and their uncertainties about the

validity of their assessments” (p. 84).

Thus, student evaluations are considered an imprecise process but are still a commonly used system. Northwestern University (1999) writes, “when we evaluate teaching, we ordinarily want to assess both what the teaching hopes to help people learn and whether it is successful with its intent” (Introduction section, ¶1). Recker and Greenwood (n.d.) write, “as we move into a new era of ‘Quality Assurance,’ universities and schools are increasingly being called upon to evaluate the quality of their courses and teaching” (Introduction section, ¶1).

This dissertation intends to examine a specific issue among the many complex issues related to the student evaluation process: the examination of both the student and the instructor self-evaluations as a means of examining this relationship. This study intends to be somewhat unique since the focus will be exclusively on comparing student evaluations with those of their course instructors (adjuncts who are part-time and full-time professors) using the same instrument. The author hopes to provide new data and subsequent insight into the field of student evaluations through the comparative nature of the study. The author also hopes the findings of the following study will contribute to a better understanding of the dual perspectives (students and professors) that are inherently involved in the SEF (student evaluation of faculty) process.

When the instructors’ feedback is also considered in addition to their students, it might lend a new perspective to the process. Bain (1982) advocates the use of the same instrument for consistency in the evaluation process by both parties.

The point seems to be that one party conducting an evaluation may overlook another party's vital perspectives. A comparative study, specifically using the same instrument, is a way to address the issue of contrasting viewpoints in the relationship between students and instructors. To actually compare and contrast students with their corresponding instructors, the author of this study suggests it is important for the evaluation forms to be the same for both the instructors and students. This is the major contribution the author intends to provide to the already large body of research on student evaluations of instructors that exclusively focuses on the student.

Other studies have looked at the importance of utilizing the same instrument when comparing student and instructor evaluations. Moses (1986) presented two primary reasons for administering self-evaluations to professors. First, she states that completing the task of self-assessment is an important "professional skill which all academics need [to learn] in their teaching" (p. 78). Second, she contends, "self-evaluation must precede self-development...focusing staff's minds clearly on the different components of teaching" (p. 78). She raises the important point of written accountability, which such an exercise elicits. Moses writes, "having committed themselves on paper, staff can more easily check their own perceptions against students' perceptions and become aware of the discrepancies which then may lead to change" (p. 78).

This dissertation will investigate the relationship between self-perception of computer science instructors in an urban setting and the perception of their adult

students. The exploration of the role of student feedback and its impact on faculty also will be analyzed here. This subject is being examined, in part, as an attempt to gain insight into the qualities that contribute to successful computer/technical teaching methods.

This brings us to another unique feature of this study, which is that it focuses on students and instructors exclusively in the proliferating field of computer science and technology. Siegel and Johnstone (1985) found that “computer studies faculty receive lower student ratings than do all other faculty on variables associated with effective teaching” (Abstract section) specifically related to their “imprecise grading policies” (p. 6), inability to “stimulate interest” (p. 6) and lack of ability to “encourage help” (p. 5) to their students. Siegel (1985) did a study that compared computer science faculty with all other faculty teaching during a term in Spring, 1984. Siegel found that “instructors in computer studies were rated lower than the other two [math and other] faculty groups” (Abstract section). This is corroborated by the Suffolk County Community College’s Office of Institutional Research (n.d.) who write “instructors of courses in the sciences appear to be rated lower than instructors of courses in the humanities” (Evaluation and Control of Central Bias section, ¶13).

From the perspective of the instructor, the content of computer science courses could pose a challenge to the new instructor unfamiliar with teaching techniques or communication skills. Such courses can cover areas like terminology, hardware, software, telecommunications, computer ethics,

applications, information management, databases and many other topics. Siegel and Johnstone (1985) write that computer science instructors “who command the knowledge and expertise to help [their students] in very real ways are perhaps weak in communicative skills” (p. 6). As stated by Cimikowski and Cook (1996), “technological changes are transforming society and the ways in which we learn” (p. 88). They add, “teachers need to be computer literate and prepared to use the computer effectively in their teaching” (p. 88). Some instructors may have years of teaching experience, but others have not taught a class before nor taken a course covering the broad field of teaching techniques. Siegel and Johnstone write, “in many cases part-time instructors need assistance to further develop the skills that are associated with effective teaching” (Abstract section).

From the students’ perspectives, they are likely to view the instructor evaluation form as the one of the rare opportunities in which they have a “voice” in the relationship. However, this adds to the challenges in this commonly used practice because it is typical that the students are the only party to conduct the evaluation with the instructor having little, if any, input in the process whatsoever.

Adding to the complexity of this process, there are topics that may need to be examined which may alter the ratings of student evaluations of instructors. Biases have been known to distort students’ ratings; for example, expected final grades, instructor gender, anonymity and the level of challenge of the class or even student retaliation. These issues have all been explored in other research papers. For example, Lawall (1998) looked at this area and found a “major area

of research has been the identification of sources of bias...[including] class size, student or instructor gender or age, time of day of class, etc.” (p. 4). In response, instructors may resort to desperate measures to appease their students. For example, Schmelkin, Spencer, and Gellman (1997) write of “stories [that] include faculty erasing forms, bringing pizza on the day of the evaluation, making pointed comments prior to the evaluation, standing over students and reading the comments out loud to the class” (p. 577). However, in a different study, Marsh and Roche (2000) conclude “teachers who want to improve their SETs [student evaluations of teachers] have far more effective and appropriate options available than resorting to counter-productive strategies such as lenient grades and light workloads” (p. 226). Marsh and Roche address this issue succinctly by stating “in contrast to popular myths, the most effective ways for teachers to get high SETs are to provide demanding and challenging materials, to facilitate student efforts to master the materials, and to encourage them to value their learning-in short, to be good teachers” (p. 226).

The students in this study are primarily working adults going to school part-time in the evenings and may require teaching skills that are unique to their needs. There is a body of knowledge that focuses on the “debate about whether distinctions should be drawn between the processes involved in educating children and those involved in educating adults” (Robinson, Arney, Munn and MacDonald, 1990, p. 1). Robinson et al. point out that until the 1970s, adult learners were treated the same as children and younger students. A researcher

by the name of Knowles “advocated the use of the term ‘andragogy’ as distinct from pedagogy, to denote the art and science of helping adults learn” (Robinson et al., p. 1). Robinson et al. (1990) note that adults require unique teaching methods and have different “study habits, learning methods and [different] motivations.” There is also the issue that “society [has] become knowledge based, and a higher education is virtually mandatory for economic success” (Hiltner & Loyland, 1998, p. 370). Hiltner and Loyland add, “education can provide [a] competitive advantage by broadening the individual’s frame of reference, perspective, and understanding, which provides increased flexibility for decision making” (p. 370) for jobs in their field. These are the students for whom pursuing advanced degrees has become important in many areas of employment.

This study is also unique in that the focus is exclusively on courses taught on a part-time and evening basis. Siegel and Johnstone write “the participation of part-time faculty in higher education is attractive, since these individuals frequently command the needed expertise and are willing and anxious to teach at night or on weekends” (Abstract section). Most of the instructors in this study (a total number of 79 of which 72 participated) are individuals who have other roles during the day (employees at other facilities, spouses, full-time parents, consultants, etc.) and teach as adjunct instructors (part-time) in the evening. Siegel and Johnstone (1985) write “as colleges and universities respond to the volume of students for courses in such fields as...computer studies, they will

continue to rely heavily on part-time faculty to teach these courses” (Abstract section). They add, “the impact of part-time faculty on higher education is particularly visible in several of the concentrations or majors where adult student interest is soaring, specifically in...computer studies” (p. 4). These working professionals have “high level(s) of expertise which they are willing to share with students” (Siegel & Johnstone, p. 4). Often these working professionals in the field of computer science view it as an honor to be asked to teach a course, given there are now numerous evening programs of study, and the assignment does not interfere with their day-time work.

The instructor evaluation instrument itself has been examined in a variety of studies including a review of the content of the items to be rated. Bain (1982) cites the following criteria which students should be concerned with when completing an instructor evaluation: “an appraisal of classroom teaching skills and the instruction of the course (materials, exams, papers), work load, course difficulty level, professor-class interaction, professor-individual student interaction, student advising, organization, displayed interest in teaching, clarity of presentation, dynamism, enthusiasm” (p. 8). But, again, where is the instructor’s voice in this process?

As stated, traditional student evaluations of instructors are just that; evaluations that exclusively involve the students’ perspective to the exclusion of the instructor in the process. This is one-sided in its approach and “obviously constitutes only one course of evaluation of teaching effectiveness” (Feldman,

1989, p. 137). This can be a very limited perspective. Feldman suggests that one response to this issue would be to compare instructors with other sources of assessment like “former students, colleagues, administrators, external observers, and the teachers themselves” (p. 137). In Feldman’s study he did just that and found “colleague and administrator ratings tend to be similar...teachers’ self-ratings and current students rating are, at best moderately similar...and the least similarity was found between teachers’ ratings of themselves and colleague ratings” (Feldman, 1989, p. 137).

The fundamental focus of this study is its comparative nature. This study could take place in any educational department, but what is unique is that both parties participated in the research without a particular focus on one party to the exclusion of the other. The Computer Science program at this urban university was of interest “owing to the rapid increase in the use of computers both in the workplace and in the home, and...[the] considerable demand for such courses” (Robinson et al., 1990, p. 2). Robinson et al.’s study, like the one presented in this paper, compared students and faculty with the goal of encouraging “discussion and reflection on the teaching methods being used, their rationale and effectiveness.”

This study is also unique because instructors are required to conduct self-assessments. The motivations for self-assessment may be varied. Moses (1986) struggles in her study (comparing instructor self-assessment with student assessments) since some instructors simply yearn for affirmation of their own

opinions. Others might be focused on their promotions and still others might merely be looking for feedback on their teaching styles (Moses). Ideally, the “primary use of self evaluation must be seen, however, in its contribution to self understanding and improvement of teaching” (p. 77).

Self-evaluation of the instructor in isolation of the student evaluation and vice versa seems to have its limitations. The author contends that the combination of both protocols will yield more relevant and informed data. Moses (1986) cites the example of a discrepancy between the two parties which will “make it more likely that staff will act on the information received from students...increasing the likelihood that improvements are made” (p. 82). Obviously, in such a study the risk exists for instructors to feel “chastised and discouraged” (Moses, p. 82). Those teachers feeling defeated by the results might “disregard them and belittle the validity” (Moses, p. 82).

It is most challenging, from the author’s perspective, to find quantitative methods that can measure a “relationship” of any kind. Many studies in the field of education employ a qualitative design technique involving interviews, transcripts and detailed interpretations. The author of this dissertation has extensive experience with such research throughout three graduate degree programs but has never before conducted an in-depth quantitative study.

The hope is that this research will lend quantitative insight into the relationship of teacher and student by interpreting numbers instead of interviews. Moses (1986) comments that “some staff seem only slowly to come to grips with getting

regular feedback from their students and reflecting in a more quantitative, comparative way on their teaching” (p. 81). Also, as stated, by examining teacher and students’ perspectives, the intention is to bring new data to one of the fastest-growing fields of concentration in education: technology and computer science.

The author hopes that other technical teachers might benefit from knowing what has worked with their predecessors and what has been perceived with positive scores. Conversely, one can always learn from previous mistakes and from a “lack of congruence” in the teacher-student relationship.

By seizing this unusual opportunity to compare instructors’ self-perceptions with their own students’ perceptions, this dissertation hopes to tease out patterns and themes that will aid in making comparisons which might lend insight into the student-instructor relationship. This dissertation intends to provide a better assessment of how the student and instructor perceive the same classroom experience using the same instrument.

In sum, the evaluation of teachers by the students is subjective and therefore not a scientific study. Retrieving the input of those being evaluated using the same format might yield additional valuable information that will enhance this teaching process and inform the system. Yet, evaluation forms like the one utilized in this study are often quantitative in nature and, thus, analyze the student-teacher relationship through numeric data. This is the attraction and challenge of this dissertation—to gain quantifiable insight into the classroom

relationship and interactions.

Statement of the Problem

It seems to be human nature for people to be curious about how others perceive them in their lives. We want to know how others honestly view our conduct, performance and level of effectiveness in any given area. There are few opportunities in one's personal life when we have access to objective, or more specifically, written data from other people regarding our behavior. Even more unusual would be the opportunity to compare our own written perceptions with those of the other party involved.

However, there are settings in life where such written information is considered commonplace and even vital, not for personal reasons, but for professional or educational reasons. For example, written performance evaluations are often required in the work setting to secure one's job or ensure professional advancement. Another example is the student who receives a final grade based on the quality of his/her course work. A third example, and the one that is obviously the focus of this study, is the classroom instructor whose students complete course evaluations at the conclusion of his/her course.

Evaluations by students are an example of one of the rare opportunities to get a glimpse into the impressions and thoughts of others, specifically how they perceive the teacher. But one must not neglect the importance of the instructor's perspective upon receiving this data as well as providing their own unique point

of view in this process. Thus, the origins of this dissertation have been established. This paper examines Student Evaluations of Teachers (SETs) from a comparative point of view with the students and their instructors using the same evaluation instrument.

How is it beneficial when the professor participates equally in the process? This dissertation study sets out to address this issue in detail. And what conclusions can be made from such data? These questions drive this study as well. Moses (1986) supports the idea of instructor participation stating, “the primary use of self-evaluation [of the instructor] must be seen...in its contribution to self-understanding [for the instructor] and improvement of teaching” (p. 77). Reid and Johnston (1999) write of the importance of selecting a “methodology [of evaluation] that [gives] weighting to both staff and student perspectives” (Aims and Methodology section, ¶2). Reid and Johnston state, “staff need to be informed by greater sensitivity to student perceptions, and that to facilitate their learning, students need to be more aware of why particular teaching techniques are preferred by their teachers” (Reid & Johnston, Conclusion section, ¶2). They conclude, “evaluation procedures need to be developed that put greater emphasis on the ambiguities that exist between what teachers and learners perceive as good teaching, both of which (of course) have a degree of legitimacy associated with them” (Reid & Johnston, Conclusion section, ¶2). Bain (1982) writes that instructor “self-assessment can be useful for improving teaching, if it’s not confused with masochistic confession on one’s faults and failures” (p. 12).

There are a variety of challenges that arise when approaching such a comparative study. For example, the sample size tested should include a large number of participants to support the study and to assist in verifying the data. This idea is supported by a researcher who used a relatively small number of participants, which in turn led to questionable research results. Ruskai (1997) writes, “the methodology of the study, which used samples of only thirteen instructors...left me skeptical about drawing any conclusions from it” (¶2). Ruskai concludes, “the findings raise serious concerns that merit further study of student evaluations in general” (¶2).

Previous studies done comparing student versus instructor evaluations have been conducted. Cashin (1988) states that “Marsh (1984) cites ten studies which correlated instructor’s self ratings with student ratings” (Validity-Instructor’s Self Ratings section, ¶1). The “correlations varied from .20 to .69, averaging .41” (Validity-Instructor’s Self Ratings section, ¶1). Cashin concluded that overall “such studies provide...support for the validity of students’ ratings” (Validity-Instructor’s Self Ratings section, ¶1). Barnett, Matthews and Jackson (2003) “compared the results of traditional student evaluations of classroom teaching with those of faculty self-evaluations” (p. 1). The results of this study, as well as other studies of a comparative nature, will be reviewed later in this paper.

Some argue that student ratings provide instructors with important feedback from the students’ view. Tang (1997) states “that students are fairly reasonable in considering important aspects of the learning process when they evaluate

professors' overall teaching effectiveness" (p. 379). However, others argue students are not capable of being adequate judges of instructor effectiveness. Feldman (1988) cites Baum and Brown (1980) where it is stated that "students do not always use appropriate criteria in evaluating their teachers" (p. 296). A conflict in perspectives might contribute to student ratings of instructors. Feldman addresses this issue writing that "students and faculty have different ideas about what is important to good teaching and effective instruction" (p. 309). The question driving this study is, what is the meaning attached to student evaluations of teachers from the instructor's, as well as the students' perspectives?

There can be conflicts of interest that impact the view of student evaluations for faculty members, particularly full-time faculty, in a university setting. For a subset of faculty in higher education there is the conflict between teaching and the pressure to "publish or perish." Some administrators feel that both research and teaching are important, "while faculty members feel that they needed to have particular strength in one or the other" area (Tang, 1997, p. 380). Tang writes "the key argument here is that teaching effectiveness is not strongly rewarded by most universities and colleges, whereas research productivity is" (p. 380). The hope is that institutions will consider "a more balanced route between instruction (teaching) and scholarship (research) to tenure university professors" (Tang, p. 386).

Clearly, from the instructor's viewpoint, the students' evaluations of teachers

(SETs) could be construed in a variety of ways depending on the meaning attached to the process by the individual and institution. There is the issue of job security for faculty in higher education as it pertains to the scoring of SETs. Gordon (n.d.) writes “student evaluations are of value to administrators and department chairs in assessing perceived effectiveness of instruction” (Student Involvement in Evaluation section, ¶2). Student evaluations are often used as a basis for “personnel decisions and faculty development recommendations in post-secondary education today” (Scriven, 1995, p. 3). Ultimately, student evaluations have a “dual purpose-as key input into personnel decisions... and an additional purpose of instructor development” (Dulz & Lyons, 2000, Introduction section, ¶2). Consequently, all this translates into instructors having strong professional and personal investments in the results.

As stated, instructors will view evaluations differently from their students if their future job security is based significantly on the results. Several researchers address this issue in their investigations. Moses (1986) states that, in higher education institutions, teacher evaluations are “one of the areas where superior or outstanding achievement must be demonstrated if academics want to be promoted” (p. 76). This is confirmed by Schmelkin et al. (1997) who write, “faculty responses indicated that [evaluations] are used by chairs and deans and the various departmental committees for reappointment, promotion, or tenure recommendations” (p. 585). Moses (1986) confirms that evaluations by students form the basis of “decisions for tenure, merit increases and promotions” (p. 77)

for higher education professors. However, Howell and Symbaluk (2001) write, “of less certainty is the extent to which student ratings beneficially affect future teaching, personnel decisions, and course selection” (p. 790).

Clearly, the stakes are high for the instructor receiving positive student ratings on their evaluation forms. In the 2003 issue of *Academe*, Gray and Bergmann continue, “at the hands of university and college administrators [evaluations are] turned into an instrument of unwarranted and unjust termination for large numbers of junior faculty and source of humiliation for many of their senior colleagues” (§1). Gray and Bergmann elaborate “administrators ...discovered they had a weapon to use against 50 percent of the faculty: they could proclaim that the half of the faculty with below-average scores in each and every department were bad teachers” (§2). Gray and Bergmann contend that, “at most, ratings may identify the very best and the very worst teachers, but they are ill designed to make fine distinctions in the vast intermediate range” (§10).

Professors may go to great lengths to receive positive ratings from students. It is suggested that “professors who want high ratings have learned that they must dumb down material, inflate grades, and keep students entertained” (Wilson, 1998, p. 1). Teachers may grade more leniently and purposely make the workloads less demanding (Wilson). Watchel (1998) writes, “faculty may tend to reduce standards and/or course workloads as a result of mandatory evaluation” (p. 194). A study conducted by Brodie (1998) attempted to “determine if students report that professors are excellent teachers when little

studying is required to receive high grades” (p. 1). The study did find that in some cases “the professor assigning the highest grades with least studying received highest evaluation” (Brodie, p. 1).

There is an argument made that student evaluations are a hindrance to teachers genuinely concerned with the quality of their teaching. According to Ruskai (1997), student evaluations “may have counter-productive effects, such as contributing to grade inflation, discouraging innovation, and deterring instructors from challenging students” (§5). Haskell (1998) writes, “it is suggested that the literature shows that student evaluations of faculty infringe on instructional responsibilities of faculty by providing a control mechanism over curricular, course content, grading, and teaching methodology” (Abstract section).

Student evaluations have been suggested to be an inhibiting factor for instructors wanting to be creative and innovative in their teaching techniques. Gray and Bergmann (2003) write “over reliance on students’ ratings also deters innovation in subject matter and methodology” (§11). Gray and Bergmann continue “an untenured faculty member can’t risk trying out a new way to teach that might improve student achievement if the faculty member knows that the old methods will produce above-average ratings” (§11). Haskell (1998) maintains, at its extreme, student evaluations produces a “pressure to comply with students’ demands [leading] directly to an infringement upon academic freedom” (Introduction section, §8). Haskell goes on writing that its “primary impact goes

to the core of academic freedom and to the quality of instruction” (Introduction section, ¶9). Ideally, instructors would get the benefit of student evaluations for the feedback for which the forms provide and for the assistance in refining their teaching practices in order to enhance the learning environment.

Reviewing and receiving the data from the student evaluation results presents challenges for the instructors as well. Once the data is gathered and processed, what if there exists a large discrepancy between the ways in which the professor views his/herself versus the way in which the students see him/her? One would venture to guess that the process would be exceedingly simple if the two parties were in constant agreement. But, what can be learned, and made productive, from such an exercise if the parties are sharply opposed to one another or similar but not identical? Moses (1986), in her study comparing students and instructors, found “both highly and poorly rated lecturers showed large discrepancies between their self-perception and student perception” (p. 76). Moses concluded from her results “that this [emphasized] the importance of using more than one source of evaluative information for decision making,” (p. 76) related to the academic future of the instructor.

This dissertation intends to explore the many issues that are raised in the complex process of involving students and instructors in the course evaluation process in a higher education setting.

Research Questions

The researcher suggests that instructors could benefit from evaluating themselves to help diminish any discrepancies between self and student perceptions utilizing identical instruments. The researcher poses three primary questions:

1. How close are instructor self-perception and student perception of instructor performance when using the same evaluation instrument? Do instructors overrate themselves compared with their students?
2. When using the same evaluation instrument, do the instructors viewed more positively (i.e., those receiving higher ratings) by their students receive higher overall ratings when compared with those instructors receiving less positive scores?
3. When there are differences between the instructor and student evaluations, are they related to the reasons that they take the course (i.e., subject major versus elective or distribution requirement)?

Hypotheses

This study is a vehicle by which instructors can compare how they perceive themselves in contrast to their students' perception using the same basis of comparison. The data can provide a connection between the two disparate parties. This sort of statistical comparative data can yield new information for professors and higher education administration responsible for making hiring and retention decisions. The researcher poses three hypotheses:

1. That when using the same evaluative instrument, the instructors will have higher scores in their self-perceptions compared to their students.
2. Those instructors with the least discrepancies from their students' ratings will have higher overall student ratings compared to the overall student ratings of those instructors with more divergent scores.
3. Those students taking the course as a requirement will be more critical of the professor than those students taking the course as an elective or a distribution requirement.

To explore these questions, this dissertation is providing the opportunity for "mutuality" (i.e., both students and instructors) in the course evaluation process in the hope that the instrument can be maximized in its usefulness and thus, its applicability. Without both parties participating "we can see that the provision of student-ratings feedback to the instructor is an incomplete tactic" (Stevens, 1987, p. 36). Stevens adds that "student ratings...[are] therefore best viewed as one means of gathering information for instructional improvement" (p. 36). But, the author suggests that the self-evaluation of the instructor is one of the "missing links" in the process. Mutual involvement is critical to the successful process as well, and the author suggests that the key element is the use of the same instrument of evaluation.

Traditionally, as discussed, students have been the sole participants in the course evaluation process. Consequently, it has been argued that the

assessment of overall teaching effectiveness was incomplete. As the process evolved, students and instructors were both involved but the instruments were not always comparable. The author suggests that equal involvement by both parties using identical instruments provides comparative data, “closes a gap” and makes for a conducive environment from which to draw more useful conclusions.

Definition of Terminology

The following are definitions of commonly used terms in this study (American Dictionary: A Random House Dictionary, Copyright © 1984):

“successful” means to accomplish or achieve

“effective” means produce intended results and be capable

“technology” means the practical application of science

“mutual” means done by two or more in relation to each other; reciprocal

“evaluation” means to appraise or to determine or fix the value of
or to determine the significance, worth, or condition of, usually by careful appraisal and study

“teach” means to impart knowledge

“quantify” is to measure an amount

“congruence” means to agree or coincide

“communication” means to make known or transmitting information

Summary and Overview

As stated, the purpose of this study is to examine the self-perceptions of computer science instructors in contrast to their students' perceptions using the same instrument. The field of computer science itself is rapidly becoming more challenging as the world becomes more complex, specialized and scientific. Though teachers in many fields have been evaluated in-depth, the computer science field is a relatively new field in education and thus, further research is required. We now live in an era of proliferating technical advances. As such, we are relying on individuals to teach future technical professionals who themselves may never have taken an education course or been trained in communication skills. The challenge for computer science instructors is to effectively deliver information that is "cutting edge" in a field that is ever changing. Keeping themselves and their students current is one of their many challenges.

Presenting technical material in an interesting and thought-provoking manner also is a challenge. Tripe (1990) writes, "in terms of [technical] teaching this is probably one of the weakest areas faced by the [technical] community...because, by and large, we...were never required to do much in the way of written or oral communication" (p. 7). Tripe observed that the typical technical teacher has taken the minimal requirements in his/her academic life of English or writing courses as well as communication or education courses.

As stated, it is even suggested in the research that technical courses are evaluated differently in general. For example, Schwarz (1997) writes "instructors who teach demanding courses, which tend to be concentrated in science,

mathematics and engineering, are often penalized with undeservedly low ratings, while teachers of easier courses are often rewarded with unfairly high ratings” (¶1). This is confirmed by Watchel (1998) who writes, “ratings in the sciences [in general] rank among the lowest” (p. 187).

The researcher’s goal is to examine the evaluation process in a fashion that will encourage faculty to reflect on their personal teaching styles in the computer science field while taking into account the perceptions their students. The goal is that the results may encourage constructive changes in individual teaching techniques. Moses (1986) warns such an exercise can involve “unwelcome surprises [for the instructor that] might close their minds toward the many positive [ratings] students” (p. 83) provide. Ultimately, the hope would be that the study would help to further enhance, strengthen and reinforce the teacher-student relationship in general, and the computer science student/instructor relationship specifically, creating an environment more favorable for learning. Moses concludes that “self evaluation and student evaluation may match and show that nothing much needs to be changed...[confirming] what we are doing and ...[increasing] our confidence” (p. 83).

This dissertation sets out to discover what defines a “good teacher” according to both the student and instructor perspective utilizing the same evaluation instrument. Similar to Reid and Johnston (1999), this study involves “subjects [who are] all mature adults [who are] often practicing professionals with” (Abstract section) widely differing experiences. Based on the standardized

university evaluation form, the two parties will be compared and contrasted to determine similarities and differences in perspectives and to draw possible conclusions to provide insight into the learning process.

CHAPTER II

LITERATURE REVIEW

Introduction

The comparison of student and teacher evaluations requires the convergence of several different research areas. The first will include the review of the history of student evaluations in general. The second area will include an examination of literature related to the reliability and validity of student evaluations. The third section will look at the literature related to the design, protocols and usages of evaluations. The final section will examine, in chronological order, previously conducted studies comparing the two populations of students versus instructors. Although there is literature on the last area, it is sparse. Literature related to student evaluations in general is more extensive.

Historical Background of Student Evaluations

In this section, a chronological and historical summary of the subject of student evaluations will be presented. To fully understand and appreciate any study, and more specifically, the one presented in this dissertation, a review of related research is essential. A historical perspective additionally provides a context from which this paper and future research suggestions can build and elaborate.

According to Wachtel (1998), “the first teacher rating scale was published in 1915” (p. 191). Seldin (1993) states “student ratings were first used in the early

1920s when students at the University of Washington were asked to fill out questionnaires about their professors” (p. A40). Thus, the earliest research seems to have begun near the beginning of the 20th century. This is supported by Wachtel: “research on student evaluations of teaching and the factors which may affect them dates back to the 1920s” (p. 191). Watchel cites a list that divides student evaluation research into four smaller time periods, as follows:

(1) 1927 to 1960 (2) the 1960’s, in which use of student evaluations was almost entirely voluntary; (3) the 1970s, which he call the ‘golden age of research on student evaluations’; and (4) the period from the early 1980s to the present day, during which time followed continued clarification and amplification of research findings (p. 192)

Clearly, the history of evaluations has evolved over time. Mason, Edwards and Roach (2002) write, “in the early years, individual instructors usually made the decision whether or not to use student evaluations, designed their own evaluation instruments, and were the only ones who saw the results” (Introduction section, ¶1). They continue, “during the 1970s, however, many universities began requiring student evaluations, standardizing evaluation instruments, and scoring the evaluation results for performance appraisal purposes” (Introduction section, ¶1). England, Hutchings and McKeachie (1996) summarize that “during the 1970s and 1980s, clear progress was made in the evaluation of teaching; student ratings of teacher effectiveness, once the exception, became the rule” (Context and Rationale section, ¶1).

In fact, the 1970s brought a new perspective to student evaluations of

teachers. Griffin and Pool (1998) write, “in the 1970s, concern often focused upon whether student rating of instruction [was] biased, or could be biased, by factors unrelated to instructional effectiveness, such as the manipulation of course grades or grading leniency” (Introduction section, ¶6). Gordon (n.d.) writes, “since the 1970’s, there has been a consensus on the purpose of student evaluations at colleges and universities” (Student Involvement in Evaluation section, ¶3). There were two categories of evaluation purposes; the summative and formative. Gordon goes on to cite Rifkin (1995) stating, “the primary purpose is formative; this is, facilitating faculty growth, development and self-improvement” (Student Involvement in Evaluation section, ¶3). This entails the use of evaluations as a tool for instructional improvement and enrichment. There are other purposes as well. Gordon writes “student evaluations are used for summative purposes and often play a vital part in tenure, promotion, reappointment, and salary decisions” (Student Involvement in Evaluation section, ¶3). Summative applications of evaluations make use of the data for administrative purposes.

Subsequently, Herbert W. Marsh was to develop the “applicability paradigm for studying the applicability of his Students’ Evaluations of Educational Quality,” otherwise known as the “SEEQ” instrument. Lawall (1998) writes that the SEEQ was developed “in the late 1970’s and unveiled in 1982 in the British Journal of Educational Psychology” (Introduction section, ¶1). Marsh is a researcher with a history of expertise in the area of student evaluations. Marsh and his colleagues

write that they:

“used SEEQ data to explore the many issues that have characterized the past decades of student ratings research: reliability, validity, and stability of results; sources of bias in the responses; the utility of ratings in administrative decisions (summative evaluation); and the usefulness of the rating for improving teaching (formative evaluation)” (Lawall, p. 1)

Marsh’s work contributed novel and innovative ideas to evaluations of instructors within the academic field. His data also contributed new insights into the student ratings placement in its historical context. Lawall (1998) writes that Marsh and his colleagues found “considerable agreement in the idea that effective teaching is comprised of a definable set of independent elements” (p. 1). Senior (1999) writes of Marsh’s SEEQ that it is “one of the most quoted questionnaire designs” (Written Questionnaires section, ¶2). The SEEQ outlined nine factors that should be examined including “Learning/Value, Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Rapport, Breadth of coverage, Examinations/Grading, Assignments/Readings, and Workload/Difficulty” (Senior, Written Questionnaires section, ¶2). Senior adds, the “number and phrasing of the questions is left open to the needs of each college” writes Senior (Written Questionnaires section, ¶2).

In the 1980s, Marsh began reviewing the validity, reliability and methodological concerns as related to student evaluations (see next section for more detail). He refers to student evaluations of teaching as “SETs” and states, “particularly in North American universities, SETs are collected almost universally” (Marsh, Hau, Chung, & Siu, 1997, p. 568). He and his colleagues

point out that SETs are generally “reliable, stable, reasonably valid against a variety of indicators of effective teaching, relatively unrelated to a wide variety of background variables, and useful to lecturers for purposes of improving teaching” (p. 568).

However, in the 1980’s, Marsh began questioning “the applicability of North American instruments” to other countries (Marsh et al., 1997, p. 69). He wondered if the data could be generalized to different environments and “noted there is danger in assuming that instruments developed in one setting can be used effectively in new settings without first testing their applicability” (Marsh et al., p. 569).

This led to Marsh’s research conducted at a university located in Hong Kong. There, Marsh and his team confirmed the SEEQ was applicable at The Chinese University of Hong Kong. Marsh, et al. (1997) write that their “results support the use of the SEEQ in this Chinese setting” (p. 568) as well as “the generalizability of findings from North America” (p. 572). The SEEQ utilized in this study was based on criteria obtained from the following sources (there were subsequent versions of the SEEQ in later studies):

“(a) [some data was] obtained from the SET literature and interviews with teachers and students, (b) students and teachers rated the importance of items, (c) teachers judged the potential usefulness of the items as a basis for feedback, and (d) open-ended student comments were examined to determine if important aspects had been excluded” (Marsh et al., p. 568)

After reviewing these results, items were selected and revised to the comprise

SEEQ version used in the Hong Kong study.

In the 1990's, research showed an increase in the use of student evaluations of instructors and their distribution becoming more commonplace. Seldin (1993) writes that "in 1993 the number of institutions using student ratings to evaluate teachers had climbed from 29 per cent to 68 percent to 86 per cent" (p. A40). Seldin notes that a Massachusetts dean "would not want to promote or tenure a faculty member without giving heavy weight to student views" (p. A40). Griffin and Pool (1998) add, "during the 1980's and early 90's, the number of published studies suggesting bias in student evaluation dropped sharply, and these were clearly outpaced by research providing evidence of the validity of student ratings" (Introduction section, ¶6).

In sum, there has clearly been an evolution in history related to student evaluations and their meaning within higher education settings. With decades of usage, it was the hope that "student evaluations represent accurate assessments of instructional quality" (Marsh et al., 1979, p. 159). What began as a voluntary process at the turn of the 20th century, has now become deeply embedded in the culture of most universities. Watchel (1998) indicated that student ratings were intended to increase the chances that teaching "will be recognized and rewarded" (p. 192). Watchel (1998) adds that the majority of research supports the belief "that student ratings are a...worthwhile means of evaluating teaching" (p. 192).

Reliability and Validity of Student Evaluations

Student evaluations cannot be reviewed in a comprehensive fashion without acknowledging the importance of the body of research related to its reliability and validity. Firstly, the concept of reliability will be addressed. An online document entitled “Student Evaluations: A critical review” (n.d.) says that a “test is said to be ‘reliable’ if it tends to give the same result when repeated” (Reliability and Validity of SEF section, ¶1). Cashin (1988) expands the definitions of reliability as an educational measurement that includes “consistency, stability and generalizability” (Reliability section, ¶1). He adds: “for student rating items, reliability is usually concerned with consistency, with interrater agreement, which varies depending upon the number of raters, i.e., the more raters, the more reliable” (Reliability section).

In addition to reliability, validity is notably documented in the research related to student evaluation of teachers. Cashin (1988) indicates that a test is said to be valid if it “measures what it is supposed to measure” (Validity in General section, ¶1). When it comes to the classroom, Cashin says, “the best criterion of effective teaching is student learning” (Validity-Student Learning section, ¶1). Examples of things that can interfere with validity are biases, more specifically, “student motivation...expected grades...and grading leniency,” (Cashin, Validity-Possible Sources of Bias-The Bad News section, ¶5).

Other researchers addressed the issue of validity related to student evaluations of instructors as well. The way in which the instructor views the results addresses the validity of the instrument. For example, Stevens (1987)

relates the issue of validity to instructor's self-evaluations when he cites the following example,

“if an instructor tends to doubt the validity of student evaluations and receives feedback that is inconsistent with his or her self-evaluation, then he or she is more likely to discount the value of student feedback and is unlikely to change” (p. 35)

Additionally, Scriven (1995) writes of one of the potential sources of validity for student evaluations is “the positive and statistically significant correlation of student ratings with learning gains” (p. 3). Griffin and Pool (1998) write that research has shown that student evaluations can give opportunities for valid measures of teaching effectiveness. Griffin and Pool observe that when instructors obtain “midterms evaluations to alter their teaching, slight improvements in instruction resulted as evidenced by the end-of-term evaluations” (Introduction section, ¶1). Overall, according to Howell and Symbaluk (2001), “student ratings are valid indicators of teaching effectiveness” (p. 790).

Reliability and validity are concepts often intertwined in general, but definitely with regard to SETs. Student Evaluations: A Critical Review (n.d.) says:

“most researchers agree that student evaluations of faculty are highly reliable, in that students tend to agree with each other in their ratings of an instructor and...are at least moderately valid, in that student ratings of course quality correlate positively with other measures of teaching effectiveness (Reliability and Validity of SEF section, ¶2)

Bain (1982) examines a variety of evaluative sources for instructors, including colleagues and previous students. He writes, “the validity and reliability

studies...clearly indicate that colleagues and students are uniquely qualified to assess certain aspects of instruction” (p. 8). New students might be less qualified than seasoned students or fellow professors to judge all aspects of teaching. Bain adds, “each source of information, student, colleague, administrator, self-assessment, offers important but limited insights” (p. 8).

The issue of reliability and validity is addressed by a variety of other researchers. Seldin (1993) writes, “hundreds of studies have determined that student ratings generally are both reliable (yielding similar result consistently) and valid (measuring what the instrument is supposed to measure)” (p. A40). Some research has found “that student ratings and comments can provide valid and reliable information that can help an evaluator determine the effectiveness of a teacher” (Northwestern University, 1999, section I, ¶1). The researchers at Northwestern University found that “student ratings are statistically reliable (i.e., they have internal stability and are consistent over time), are more statistically valid than are colleague ratings, and are not easily or automatically manipulated by grades” (Northwestern University, section I, ¶1). In fact, some intellectually challenging courses average higher ratings than easier courses with light workloads. Most importantly, it has been suggested that student ratings work to tell if the instructor has “reached” their students.

Student biases have been shown, in research related to evaluations, to interfere with the reliability and validity of the results. Seldin (1993) addresses this issue stating, “some faculty members and administrators argue that factors

beyond professors' control bias the ratings" (p. A 40). Biases cited include age of students or even the gender of instructor. In an extreme example, Riger (1993) cites an email that said, "students expected female faculty members to be warm and nurturing, and when the faculty were perceived as providing 'inadequate levels' of warmth and nurturance, their ratings took a dive" (Sharyl Bender Peterson/The Colorado College section, ¶1). However, some of the other variables that may bias results are class size, time of day of class, grading patterns or the attractiveness of the instructor. In another example, Miller (2003) cites "a recent study [where] two researchers at the University of Texas at Austin concluded that more attractive professors outscored their more homely peers on teaching evaluations" (Article Preview section, ¶1). Even more surprising was a paper that cited students whose biases included "comments on a professor's clothing, hairstyle, or personal hygiene" (Wilson, 1998).

Reliability and validity has received a significant amount of attention in the research related to student evaluations of instructors. With this in mind, according to Stevens (1987), student evaluations in general "provide reliable" (p. 33) methodology for obtaining data regarding instructors. Stevens states, "reviews of the validity of student ratings have tended to support their usefulness as a measure of instructional effectiveness" (p. 33). He says, "student ratings instruments...provide the most reliable and cost-effective means of obtaining feedback" (p. 36).

However, student evaluations of instructors are often a dreaded exercise for

the faculty specifically. According Olp, Watson and Valek (1991) “faculty appraisal is often viewed with as much enthusiasm at a trip to the dentist” (p. 3). They note “the motivating factors [for evaluations] are achievement, recognition, responsibility and advancement” (p. 3). As previously, stated, Marsh and Roche (1993) add “SETs are widely used also for personnel decisions” (p. 218). Future job security for instructors can impact the meaning of the results, and, therefore, “the reliability and validity of student ratings have been a source of discontent for faculty awaiting tenure decisions” (Kaufman, 1981, p. 2).

The forms and the scoring processes related to instructor evaluations are vulnerable to problems of reliability and validity. By closely examining the variables that comprise an evaluation form and looking at how to meet the objective of the instrument, the institution attempts to create a tool that is reliable and valid. This ideally would translate into fair and reasonable scoring as well. Seldin (1993) gives the example, however, of minimal differences between faculty members’ scores, which sometimes attract attention. He writes that “a professor who receives a rating of 3.72 is not a significantly better teacher than a colleague who receives a rating 3.70” (p. A40) though some administrators might still examine these differences. Seldin continues by stating, “even a carefully developed student rating form can be invalidated by poor administration, including a sporadic rating schedule or instructions that bias responses” (A40). Seldin believes the questions themselves should be designed to measure designated areas of interest. He states “if the purpose is improved teaching, the

form should include...diagnostic questions [regarding] specific teaching behaviors” (p. A40). If the form is used for personnel decisions (i.e. tenure), questions should focus on the “faculty member’s performance” (Seldin, p. A40).

Some instructors view their students as the opposition party in the evaluation process. The fact is that students’ opinions can be considered impaired when viewed by some, and that fact can interfere with reliability and validity issues related to course evaluations. Northwestern University (1999) supports this idea, stating that students “are not always well equipped to judge the course as an intellectual product, to determine whether it is appropriate to the curriculum or sufficiently rigorous” (section III, ¶5). Students’ opinions may be challenged as being a good resource: “student ratings can provide valuable information, but they [the students] cannot always tell evaluators everything needed to make valid, reliable assessments of teaching effectiveness” (Northwestern University, section III, ¶5). A newsletter from The Center for Teaching and Learning (1994, p. 1) says, “one objection to student ratings is that they are not valid measure of teaching effectiveness; that students are not able to assess good teaching and therefore evaluations represent nothing more than a popularity contest.”

Researchers argue that students may not “have enough content knowledge to effectively evaluate teaching” (Barnett et al., 2003, p. 1). And, students may not have the ability to make decisions regarding faculty or are not qualified to have their opinions valued so greatly (Schmelkin et al., 1997). Lawall (1998) addresses this issue directly: “students are too immature, capricious, and

inexperienced to give reliable feedback on teaching” (SEEQ Research section, ¶11).

Others argue that students are capable of evaluation. Bain (1982) writes, “research demonstrates that students offer reliable and valid assessments when asked appropriate questions” (p. 7). For example, Howard, Conway and Maxwell (1985) contend that “student raters typically have more than 20 times the exposure to their instructor’s teaching at the time they make their judgments than do colleague and trainer observer raters” (p.195). Students’ insights, combined with their in-depth experience with the instructor in the classroom, can boost their credibility as constructive observers of instructors.

Some disagreement exists about the value of the evaluation process as a whole. According to Stevens (1987), student evaluations in general provide valid methodology for obtaining data regarding instructors. Stevens states, “reviews of the validity of student ratings have tended to support their usefulness as a measure of instructional effectiveness” (p. 33). Other researchers maintain that student evaluations can confuse the instructor. Stevens contends, “the effectiveness of providing feedback for instructional improvement is dependent on conditions that allow feedback information to be received favorably, and, once received, to be applied as part of a meaningful strategy for change” (p. 34). Clearly, the institution’s culture and attitudes regarding SETs plays a central role in how the information is processed and received by the instructors.

Because there is not any “universal criterion” (Marsh, 1982, p. 264) used for

evaluating effective teachers, many types of student evaluations may exist and may be used to rate instructors. Consequently, it is difficult to measure the student-teacher relationship in any consistent way. Howard et al. (1985) state one can only assume “the availability of some accepted method or methods for the accurate assessment” (p. 187) of higher education teaching effectiveness. In fact, in their study, they explored the perceptions of former and current students. They found that “former-student and student ratings evidence substantially greater validity coefficients of teaching effectiveness than do self-report, colleague, and trained observer ratings” (p. 195).

Design, Protocols and Usages of Evaluations

It has been established that evaluations of teachers are common practice at most education institutions. Marsh and Roche (1993) state that student evaluations of teachers “are commonly collected and frequently studied” in the United States (p. 219). We know how widespread the use of evaluations is when Schmelkin et al. (1997) write “student ratings of instruction can be found in some form or another at most American colleges and universities” (p. 575).

It seems important that the design, protocols and specific uses associated with teaching evaluations be discussed at this point in this study. El-Hassan (1995), who cites “Gage (1972) as cited in Marsh 1984,” (p. 411) writes the following outline of the main uses and applications for student ratings of instructors:

(a) diagnostic feedback to faculty about the effectiveness of their teaching, (b) a measure of teaching effectiveness to be used in tenure/promotion decisions, (c) information for students to use in the selection of courses and instructors, and (d) an outcome on a process description for research or teaching; that is, they could answer questions like How do teachers behave? Why do they behave as they do? And what are the effects of their behavior (p. 411)

There are clearly multi-dimensional ways in which SETs can be utilized.

To review the specific usages and applications of evaluations, the design of the form itself needs specific attention, and the research reflects this idea. Some of the suggestions for recreating evaluation forms include having faculty themselves have direct input into the evaluation system. This notion leads to the idea that “departments and schools can then take responsibility for developing their own evaluation methods and evaluation criteria” (University of Michigan, 2004, Some Principles of Teaching Evaluation section, ¶2). One important dimension to consider is the specific type/area of discipline which may “require different methods and setting for instruction” (University of Michigan, Some Principles of Teaching Evaluation section, ¶2). In higher education settings, the evaluation construction could vary greatly depending on the course content. For example, teaching methods may include “lecture, discussion, lab, case study, small group interaction, studio, field work, clinical work, etc” (University of Michigan, Some Principles of Teaching Evaluation section, ¶2) depending on the subject matter and field of study. This variety in course structure might need to be reflected in the design of the items on the evaluation forms to increase their

relevancy.

The literature suggests that the wording and construction of the form itself must be carefully scrutinized. The Center for Teaching and Learning (1994) states, “the questions on the student evaluation forms should correspond to the aspects of teaching that the department considers important” (p. 2). In addition, the form should not be too lengthy; a maximum of “twenty-five items” is suggested (p. 2). The specific design of the rating scale partially determines the usefulness of the data. The Center for Teaching and Learning (1994) says the “nature of the rating scale will play a role in how useful the data will be” (p. 2). And then there is the option of a “comments” section for students, which offers an open-ended narrative opportunity. The exact wording of the items has also been addressed. The Ad-hoc Committee on Student Evaluations of Ramapo College of New Jersey (2001) “recommends that only a narrative, qualitative form be used to accumulate student satisfaction data” (Introduction section, ¶1). In their opinion, the Ad-hoc Committee concludes “that a numerical student evaluation form provides potentially inaccurate and misleading information, of little value in assessing teaching effectiveness” (Introduction section, ¶1). Wide variability exists in the design and format of each evaluation form.

The protocol followed throughout the process of the evaluation process requires attention as well. Exclusively administering the evaluations at the end of a course is the most common way instructors are evaluated (Barnett et al., 2003). The Center for Teaching and Learning (1994) addresses the alternative

procedures to be followed when administering the forms to students. They suggest giving the form to students “a week or two before the end of the semester” (p. 2) rather than near the final exam since students might be distressed at the later time. In 1997, Marsh and Roche emphasized the importance of “SET feedback and consultation” (p. 1194) through the use of teacher self-evaluations including both mid-semester and end-of-the semester evaluations by students. As stated, this “feedback-consultation intervention” (p. 1194) was a contrast to the traditional end-of-term only opportunity for student input. The result was that these interventions were concluded to be an effective process for improving teaching effectiveness (Moses, 1986).

Student evaluations of instructors can include a variety of factors as part of the protocol: they should be anonymous, the students should have ample time to finish the form completely, someone should oversee the process and answer questions, and the instructor should exit the room during the evaluation process (The Center for Teaching and Learning, 1994). It is mandated that “someone other than the instructor should distribute and collect the forms” (Center for Teaching and Learning, p. 3).

Another aspect of the protocol is the timeliness in which the results of the evaluations are returned to the instructors. One way to increase the relevance of the evaluations is to have the results returned to the instructor in a more rapid manner. As Marincovich writes (1998), “if teaching evaluations data are to be taken seriously by faculty... [they must] receive their results as close to the

administration of the forms as possible” (p. 4). An instructor may, in some cases, receive the data long after the course is completed. Coburn (1984) notes “if the results of student ratings are not reported in a timely manner, their usefulness can be compromised” (Reporting the Results section, ¶1).

One cannot look at the student evaluation protocols without looking at the anonymity factor for the students completing the forms. This feedback is primarily retrieved from students who are secure in the knowledge that their thoughts are anonymous. This, in turn, might encourage a type of honesty, which might not occur otherwise. This is confirmed by Riker and Greenwood (n.d.) when they write “research reports somewhat lower ratings when student responses are anonymous, especially if evaluations are administered before grade assignments are made” (Administrator Perspective section, ¶1). If their names were required on evaluations, students might be more likely to screen their thoughts. This may stem from the perceived fear of retaliation in their grading or intimidation intrinsic in the student-teacher hierarchical relationship. On the contrary, Fries and McNinch (2003) addressed the issue of anonymity related to student evaluations and cited research that “looked at having students sign the forms” (p. 333). They concluded that “the rationale behind such a suggestion seems clear: by introducing a measure of personal accountability into the evaluation process, students hopefully will be encouraged to take the process of evaluating their instructors more seriously” (p. 333). They continue, “our sample shows that if students have something negative to say on evaluations,

they tend not to sign their forms even when asked to do so” (p. 339). They observed that, “asking students to sign the evaluation forms leads to more positive ratings across the board in all categories” (p. 341) of teaching.

Protocol also includes addressing the question of what populations of individuals will have access to the results of the evaluations. Coburn writes “one of the most important decisions to be made is who will see or use the results” (Reporting the Results section, ¶2) of student evaluations of instructors. Students clearly have a vested interest in course evaluations. One use of student evaluations results: they “can be used by other students to select courses and instructors” (Coburn, 1984, Arguments in Support of Student Ratings section, ¶5). Some schools report the results in “student newspaper or student published books” (Coburn, Reporting the Results section, ¶3). All who use the ratings must be careful to avoid placing inappropriate emphasis on student responses on evaluations. It should be noted, “student ratings are but one component of a comprehensive instruction evaluation system” (Coburn, Reporting the Results section, ¶4). The suggestion is to keep the student evaluations of instructors’ results within a broader context and not rely on them to the exclusion of other components of teaching. Coburn suggests that, most importantly, “student ratings [should be used to] encourage communication between students and their instructor” (Arguments in Support of Student Ratings section, ¶4).

Instructor evaluations play one part in a host of components related to the

student-teacher relationship. Palmer (1990) writes that the whole process of teaching students is “an act of generosity...and is always risky business” (p. 11). He does not advocate the use of “evaluations that are collected on questionnaires and published” (p. 16). Instead, Palmer believes that evaluations of instructors should take place “publicly at the end of every second or third class [involving] a time of open reflection on how things are going” (p. 16). The evaluation process can be demystified “when a class knows that it will be asked periodically to assess its own progress, everyone-the teacher included-comes to class with more intention and wit, more sense of being in this together” (Palmer, p. 16).

Accountability to the university administration regarding teacher effectiveness holds significant importance in the current academic setting and directly relates to the usage of instructor evaluations (Reid & Johnston, 1999). Student ratings “provide information regarding teaching effectiveness to faculty for the purpose of feedback and improvement, to administrators for the purpose of faculty promotion, and to other students for the purpose of course and instructor selection” (Howell & Symbaluk, 2001, p. 790). As previously stated, Marsh and Roche (1993) add “SETs are widely used also for personnel decisions” (p. 218).

One risk related to the usage of instructor evaluations is that they will be a source of discouragement. Discrepancies can exist between students and professors when it relates to evaluations. Take the example of an instructor who may be confused when reading the class evaluations perceiving his/her

knowledge base as a source of pride only to be faced with low scores. Instructor morale may be affected if the individual perceives him/herself in stark contrast to that of his/her class. In such cases, Seldin (1993) suggests that the student ratings can “lead to anxiety, discouragement and diminished enthusiasm for teaching” (p. A40). On the other hand, a new instructor is likely to feel affirmed and motivated by positive student perceptions after feeling insecure during the actual classroom experience. Student evaluations may be the one chance an instructor will have to know exactly what the students’ think of their classroom performance.

Questions about the use of these often subjective instruments will continue to exist. Teachers, administrators and researchers constantly ask themselves, “what in fact do evaluations really measure?” Are the results accurate, and by whose standards? The fairness of such tools are controversial, especially when they are used for making significant life decisions of faculty based on students who may focus on emotional issues rather than skill level or other more objective variables. Discontent has been voiced by instructors who feel evaluations are “(a) invalid, (b) unreliable, (c) highly correlated with grades, and (d) popularity contests” (Schmelkin et al., 1997, p. 576).

Clearly, the design, protocols and usages of student evaluations of instruction are complex. The literature reflects the many diverse and controversial issues that result from the often quantitative instruments used to measure the student/instructor relationship.

Comparative Studies of Instructor and Student Evaluations

While numerous studies review student evaluations in general, the author found a dearth of studies specifically comparing student and instructor evaluations. Of those studies found, a description of each will follow, in chronological order, to explore the evolution of this specific body of research over the last fifty years.

As far back as 1955, Yourglich searched for the “ideal” teacher and “ideal” student by comparing their perspectives. Her study explored the “correlation” of concepts that comprise a model learning relationship between teacher and students. The same questionnaire was given to 35 teachers and 101 college students from a variety of academic departments. This study looked at the correlation of agreement and disagreement in the rankings of the students and instructors on an evaluation instrument. The following were the results of Yourglich’s study; “teachers and students seem to be more in agreement as to what an ‘ideal student’ is, and less in agreement as to what an ‘ideal-teacher’ is” (p. 63).

In 1973, Centra compared college teachers’ self-ratings with the ratings given by their students. The sample consisted of 343 faculty members and their respective students. Centra found “teacher self-ratings had only a modest relationship with the ratings given by students (a median correlation of .21 for the 21 items)” (p. 287). There were not any differences related to the following

variables: gender of the instructor or number of years of previous instruction experience.

Centra's (1973) research pointed to the need for a comparative study as a "source of information for performance improvement and, to a lesser, extent, as input into performance evaluation" (p. 287) for instructors. He argues that student evaluations should not be used for a basis "for decisions on promotion or salary [since they] are not likely to have much validity" (p. 287). He states, "discrepancies between self-ratings or self-descriptions and those provided by students would underscore the need for student feedback to the instructor as well as highlight specific areas of instruction where feedback is most essential" (p. 287). Centra's data "disclosed a modest relationship between the two evaluations" (p. 293). Centra notes "in addition to the general lack of agreement between the self and student evaluations, there was also a tendency for teachers as a group to give themselves higher ratings than their students did" (p. 287). Centra theorizes that "this tendency might be viewed as 'human,' or certainly not surprising...since people do not see themselves as others see them; teachers and the way they see their instruction is apparently no exception" (p. 293).

In 1974, Sagen looked at student and faculty self-ratings but added in the perspective of the department chairperson in a small liberal arts college. In addition to looking at overall instructor ratings, Sagen's study examined "what aspects of instruction are considered most important by each group" (p. 265). The result of this study revealed "little agreement" amongst all parties related to

instructor ratings and “at best modest agreement between students and department chairmen concerning certain specific aspects of instruction” (p. 265). Overall agreement related to instructional effectiveness was minimal between “student ratings, faculty self-ratings, and department chairmen” (p. 265) ratings. Sagen concluded, “students, faculty, and department chairmen do not arrive at ratings of overall effectiveness in the same manner” (p. 271). Sagen notes that “students seem to stress the instructor’s ability to facilitate mastery of the subject, whereas faculty and department chairmen place more emphasis on personal qualities of the teacher” (p. 271). His research leads to the conclusion that “no single measure [of instructor evaluation] is clearly superior” (p. 271). A better method for evaluation “is to employ several measures and to base a final judgment on the consistency among instruments or upon a composite of measures employed” (p. 271). He argues that “the result of this...study of instructional effectiveness...lead to one basic conclusion: that faculty evaluation should be treated realistically as the appraisal of an exceedingly complex professional performance about which we still know relatively little” (p. 272).

In 1979, Marsh et al. conducted a study where “faculty evaluated their own teaching and were evaluated by their students in each of two courses” (p. 149). They note, “both faculty and students used essentially the same evaluation form” (p. 158). Despite the skepticism of the faculty “regarding the validity of student ratings, there was considerable student-faculty agreement in the ratings obtained” (p. 159). They concluded, “these findings reaffirm the validity of

student evaluations, suggest the possible usefulness of faculty self-evaluations, and should help reassure faculty about the accuracy of the student ratings” (p. 149). The pressing concern among faculty, according to Marsh et al., “is whether or not these ratings—often the only measure of teaching effectiveness regularly available—actually reflect effective teaching” (p. 149). Marsh et al. present the challenges of previous studies including that of being “limited to a specialized setting or [employing] criteria that are open to criticism” (p. 149). Although such variables as “class size, reason for taking the course, workload, and grade point average have little relationship to such ratings,” (p. 149) the validity of such ratings is still questioned in this study.

In 1982, Marsh had 329 college instructors evaluate their own teaching “with the same 35-item rating form that was used by their students” (p. 264). Marsh wrote, “these findings demonstrate student-instructor agreement on evaluations of teaching effectiveness [and] support the validity of student ratings for both graduate and undergraduate courses” (p. 264). Marsh reports “before the potential usefulness of student ratings can be realized, faculty...have to be convinced that student ratings are valid and relatively free of bias” (p. 266). He adds, “in spite of faculty skepticism concerning the validity of student ratings and their belief that many sources of potential bias substantially affect the ratings, there was good student-instructor agreement” (p. 277). His study provided evidence that “student ratings show good agreement with instructor self evaluations of teaching effectiveness” (p. 278).

Also in 1982, Bain noted that student “ratings substantially different from an instructor’s self-assessment may provide sufficient motivation to change” (p. 7). He qualifies that idea by saying that the instructor must possess the motivation to change and the guidance to assist in the process. One of the ways instructors have been evaluated in addition to their students is by colleagues and/or school administrators (Bain). Also, instructors have been videotaped and then “receive immediate reinforcement and suggestion for improvement from colleagues or consultants” (Bain, p. 7).

In 1988, Feldman analyzed previous studies regarding “the extent to which students and faculty...differ in the criteria each group uses in evaluating teaching” (p. 296). The studies specifically examined were those “in which both students and faculty at the same school or schools were asked to indicate the instructional characteristics they considered important to good teaching” (Feldman, p. 296). Feldman’s findings show “faculty members...not to be much different from students in their views on good teaching” (p. 309). From the studies Feldman reviewed, it was “clear that students and faculty were similar in placing high importance on teachers being prepared and organized, clear and understandable, and sensitive to class level and progress” (p. 311). Both parties also valued “instructor enthusiasm,” instructor “knowledge of the subject matter,” “instructor’s fairness and impartiality of evaluation,” and “friendliness” (p. 311). Faculty and students alike gave low importance to “clarity of course objectives and requirements” (Feldman, p. 311). Differences emerged between

students and faculty related to “the importance of the teacher stimulating their interest in the course and in its subject matter” (Feldman, p. 312) with students putting greater value on this idea than the instructors. Students “placed moderate importance on teachers’ elocutionary skills [while] faculty felt this instructional aspect to be of low importance” (p. 312). Another difference was that “students also placed low importance on teachers setting high standards of performance and motivating students to do their best as well as on encouraging self-initiated learning, whereas faculty rated these aspects of teaching as moderate in importance” (p. 312). Feldman concludes “the fact that certain similarities and differences in the criteria students and faculty use in determining good teaching can be found across studies creates some confidence in their existence” (p. 313). However, Feldman adds, “any generalizations based on these particular comparisons are tentative, at best” (p. 313). In agreement with this dissertation study, Feldman writes “an obvious need thus exists for future research in which the data on the views of students and faculty and the data on the actual specific and overall student rating of faculty are collected from matching samples” (p. 313).

In 1999, another study compared medical students and residents with their corresponding faculty upon viewing two videotaped medical school lectures. Leamon, Servis, Canning and Searles (1999) observed “surprisingly little investigation comparing student evaluations with faculty peer evaluations of teaching in preclinical medical school courses” (p. S22). This study’s “most

salient finding [was] the absence of differences between student ratings of medical school lecturers' effectiveness and faculty ratings of the same lecturers" (Leamon et al., p. S23). They add that their findings, "based in a medical school setting, mirror the majority of previous studies based in college and university settings that compared faculty and student ratings" (Leamon et al., S. 24).

Also in 1999, in a study by Reid and Johnston, the aim was "to provide evidence to what both staff and students consider to be elements of good teaching and to inform appropriate staff development in an attempt to improve teaching effectiveness" (Aims and Methodology section, ¶1). They raise the issue of differing views between the two parties, writing that "student and lecturer perceptions of what is required do not always coincide" (Reid & Johnston, Conclusion section, ¶2). Their study supports the author's idea that mutual "participation be extended beyond the traditional teaching boundary" and "become an essential element of adult learning" (Reid & Johnston, Conclusion section, ¶2).

In 2001, a study took an international perspective when comparing student and instructor evaluations comparing universities in the United States, South Africa and China. Miller, Dzindolet, Weinstein and Xie (2001) write the "similarity of faculty and students' views of teaching effectiveness has become an important issue in light of the widespread use of student evaluations" (p. 138). They conclude, "the value of using student evaluations to measure teaching effectiveness is directly related to the similarity between faculty and students'

conceptions of effective teaching” (Miller et al., p. 139). The issue of getting high ratings is sometimes believed to go to those instructors who give high grades or who are entertaining. Miller et al. also point to cross-cultural differences amongst student ratings. For example, in Thailand students “rated teaching competence and motivation skills as most important” and in Spain “students rated teaching competence and motivation skills as most important” (Miller et al., p. 139). These researchers conclude,

“a high degree of similarity in views of teaching effectiveness was found between instructors and students [among instructors and students from other countries] on items concerning preparation, evaluation, presentation, and opportunities for student inquiry, suggesting that instructors and students use the same criteria for rating teaching effectiveness” (Miller et al., p. 138)

Miller et al. write, “similarity among faculty and student views concerning effective teaching suggests that students use reasonable criteria to rate their instructors” and supports the “usefulness” (p. 141) of this mutual process of evaluation.

Bosshardt and Watts exclusively focused their 2001 study on comparing student and instructor evaluations in Departments of Economics. They write, “most economics departments use end-of-term student evaluations of teaching, but the relationship between instructors’ assessments of their own teaching and their students’ assessments is unknown” (Bosshardt & Watts, p. 3). The study revealed “important differences between instructors’ and students’ perceptions of what constitutes good teaching” (Bosshardt & Watts, p. 4). They illustrate the fact that “few studies have directly estimated the weights for individual SET items

on overall student evaluations of instructors and none compares these results to weights calculated from instructors' self-evaluations" (Bosshardt & Watts, p. 5). A strength of their study is that "instructors are asked the same questions in a self-evaluation [and] the instructor and student weights can be compared" (Bosshardt & Watts, p. 5). They add the "primary value of such comparisons may be to see whether instructors and their students perceive the same strengths and weaknesses in instructors' teaching" (p. 5) and found "instructors and students have different views on the relationships between overall instructor effectiveness" (Bosshardt & Watts, p. 13). Overall, however, their research indicated a positive correlation between perceptions of both instructor and student for teacher effectiveness.

Another 2001 study further explored the value of comparing student and faculty perspectives as they pertain to the publishing of course evaluations results. Howell and Symbaluk (2001) specifically found that students preferred "published ratings" while "faculty cited numerous disadvantages of published ratings and rated the likelihood of potential costs as high relative to students" (p. 790).

In sum, there has been research of a comparative nature conducted on student versus instructor ratings as related to the evaluation process in higher education settings, but the need appears to exist for future research. The author hopes this study will add new information to the current body of literature.

Summary of the Relevant Research

This dissertation focuses specifically on a master's degree level of instruction in higher education and the evaluation of the effectiveness of a mutual evaluation process. Such an advanced level of teaching would presume a commitment to adult learners that focuses on educators who "should be especially interested in learning about the quality of their teaching" (Olp et al., 1991, p. 3). They suggest that evaluation results should be provided "in a manner that is supportive rather than threatening" (p. 3). Olp et al. further discusses the need for an appraisal system as one "focused on the essential mission [of] teaching and learning" (p. 8). This chapter shows, however, that evaluations are "not always a harmonious venture" (Olp et al., p. 8).

There is a large body of research on SETs that focus on "their validity, reliability, relationship to other variables, and potential biasing factors" (Schmelkin et al., 1997, p. 575). Perspectives of the various groups directly affected by evaluation, especially the faculty members, are not equally represented. Schmelkin et al. write "comparatively little research attention has been devoted to the perspectives of the different groups involved including...the faculty who are being rated" (p. 576). Schmelkin et al. agree that there is "a distinct need to conduct empirical studies that assess faculty's opinions about the usefulness of [the] ratings" (p. 577). Gould (1991) writes that "in short, self-evaluation [by instructors] ...can be a particularly effective means of getting teachers to confront discrepancies between self-perceptions and the perceptions

of others, especially students” (p. 11).

CHAPTER III: METHODOLOGY

Introduction

This chapter is organized into six sections including: the sample population studied, the instrumentation employed, the data collection procedures, the research design, the treatment of the data and the limitations of the study methodology.

Sample Population to be Studied

The study was conducted at a private, urban university in the northeastern United States in the adult part-time college within the computer science department, with an enrollment of approximately 3,000 adult students. Of the students who completed the section of the form indicating the college in which they were enrolled, there were 1,324 students, with 1,218 enrolled in MET Computer Science, 23 in Sargent College (SAR), 2 in College of Basic Studies (CBS), 26 in College of Liberal Arts (CLA), 20 in College of Communications (COM), 18 in School of Engineering (ENG), 1 in School of Education (SED), 1 in School of Fine Arts (SFA) and 15 in School of Management (SMG) and there were not any students enrolled in "other" departments.

The 72 instructors who participated in the study included adjunct and full-time professors. A total of 1,250 (out of 1,324) students responded to the "Reason for Taking the Course" query on the course evaluation form. The students who took

the class for a “Major Distribution Requirement” for the Computer Science program totaled 773. A total of 109 students took the course to fulfill a “Distribution Requirement.” Those who took the class as an “Elective” course totaled 368.

Instrumentation

Attached is a copy of the University “Course Evaluation Form” (see Appendix A: Course Evaluation Form) that was used in this study. The evaluation was a “Likert Scale” standardized form consisting of 16 items ranging from “the instructor is well prepared for class” to “overall I rate this instructor as a good teacher.” The upper third of the forms were kept anonymous, requiring the student only to provide the first five letters of the instructor’s last name, the course number, the semester/year, the college in which the student was registered and the reason for taking the course (i.e., required course, elective class or distribution requirement). Comments were “optional” and space was provided on the backside of the form, but those items were not rated according to their substance. The forms were used to evaluate the entire faculty at the end of the spring and fall semesters. The forms were then sent to a centralized office at the University for tabulation of the scores. The data were then returned to the department chairpersons with the results, in the form of graphs.

Students evaluated the faculty in response to “statements” written on the lower portion of the form. They were asked to “use a #2 pencil in dark marks” to

indicate their choice on the Likert rating scale. All sixteen statements are followed by (1) strongly disagree, (2) disagree, (3) sometimes agree, (4) agree, (5) strongly agree and (NA) not applicable. Faculty also provided their self-evaluations using the same instrument.

Data Collection Procedures

Permission was obtained in advance from the Computer Science Department Chairman and the Dean of the College to begin collecting the data. Initially, the faculty was contacted by memo and/or by phone to explain the study and allow some time for their consideration of whether they wanted to participate. This was followed by the distribution of the “informed consent form” (see Appendix B: Informed Consent Form) for their signature. The letter of “informed consent” was distributed to all faculty members in the department over a period of two semesters to indicate their willingness to participate in the study. Research subjects (i.e., the instructors) were recruited on a volunteer basis. Once the signatures were secured on the letter of consent, the data were collected and reviewed by the researcher. To ensure confidentiality, each instructor and student was assigned a number that kept all identities anonymous.

Before the end of each course, a cover letter was sent along with the course evaluations that outlined the protocol for disseminating the evaluations (i.e., the instructor not being in the room at the time of the evaluation and a neutral third party collecting the papers and returning them to the main office of the college).

This procedure was consistent with that of Marincovich (1998) who suggested that the University Registrar's departments should disseminate the evaluation forms to the individual departments accompanied by "a letter...from the academic dean [to]... situate them in an academic/scholarly context" (p. 4).

Full access to the course evaluations of the Computer Science instructors and students was granted to the researcher for the Spring 1996 and Fall 1996 semesters. For these two semesters, in addition to the students, the faculty also was invited to complete self-evaluations on the same forms. This invitation was extended to the professors through a memo signed by the Dean of College (see Appendix C: Approval Memo Signed by the College Dean). The intention was to glean faculty attitudes toward their own teaching skills compared with their students. The intention was also to observe if there were any discernible patterns of overall instructor self-perception compared to the student perceptions.

Data were collected from 1,452 graduate students in the Computer Science Master's program. The students rated 79 adjunct and full-time instructors. Some of the students and faculty had taken or taught more than one course and their data is included in the study. Some of the professors taught more than one section of the same course and their data is included as well. Finally, some instructors did not complete the self-evaluations (in seven of the 79 courses offerings) although their students did. In this case, the students' responses were still included in this study.

As indicated on the consent form, in addition to the assurance of complete anonymity, all participants in the study were offered the option to receive the results at the conclusion of the project. Finally, the participants were given the option to withdraw from the study at any time during the process without explanation and one instructor took advantage of that option.

Research Design

To review the hypotheses questions that drove the study, the following are restated:

1. The author hypothesizes that, using the same instrument, the instructors will have higher scores in their self-perceptions compared to their students.
2. Those instructors with the least discrepancies from their students' ratings will have higher overall student ratings compared to the overall student ratings of those instructors with more divergent scores.
3. Those students taking the course as a requirement will be more critical of the professor than those students taking the course as an elective or distribution requirement.

For hypothesis number one, the author compared the professor versus the students' responses. The author looked for statistically higher ratings by instructors compared to their respective students (based on the means and standard deviations of each population) for each of the 16 items. For hypothesis

number two, the author compared the mean instructor and student scores for each item and separated them into two groups; those items that were statistically similar and those that were statistically different. The student scores for each group were then compared to determine if the student scores in the statistically similar group were significantly greater than those in the group that was statistically different. For hypothesis number three, the author compared the mean instructor and student scores for each item to determine if the scores of students required to take the course for their major was significantly lower than those of students who were taking the course as an elective or distribution requirement.

Treatment of Data

The raw data was entered into Excel (MicroSoft® Excel 97) spreadsheets for each of the participants. The individual responses to each of the 16 items, plus the demographic data located at the top of form (i.e., professors' last names, course number, semester/year, college registered and reason for taking the course) were entered for all students and faculty members. Each student was assigned a unique number starting with number "1" and ending with number "1,452." Each instructor was assigned their own individual number commencing with "1,500" and ending with "1,578." Each student's and each instructor's responses to every variable was documented with his/her assigned rating (see Graph Section).

The data in the Excel (MicroSoft® Excel 97) database was evaluated utilizing basic statistical analyses (e.g., means, standard deviations, median, range and Student's t-test). More sophisticated statistical analysis was performed using Microsoft Windows SAS/STAT® Software Version 8.1 (e.g., Clustered Data Analysis and Discrepancy Analysis).

Overall Instructor Versus Student Rating

Using Excel, the mean and standard deviation of all the students and professors, respectively, was calculated for each variable (1-16). This was justified because, although the Likert Scale responses were not normally distributed, "when the sample become very large, then the sample means will follow the normal distribution even if the respective variable is not normally distributed in the population, or is not measured very well" [StatSoft (StatSoft, Inc., 2001, Version 8.1), When to Use Which Method section]. Values of $p < 0.05$ (2-tailed Student's t-test with unequal variances) were considered statistically significant.

The initial statistical analysis utilized the total number of participants in each group. This approach assumed no correlations within and between classes, which is probably unwarranted since student responses in the same classroom are likely to be more similar to each other than the student body as a whole.

Instructor Versus Average Student Rating

Responses from students rating the same class will be more similar to each other than two randomly chosen student ratings. Therefore, ratings from students within the same class are not independent and this issue was addressed in the analysis. One approach involved taking the average of the student responses for each class and each question (i.e., reducing the student data to one number) and then comparing this average student rating to that of the course instructor. The instructor self-rating was compared to the average student rating *for each of the 16 questions* and the *sum of the 16 questions* via a paired t-test. The difference between the instructor rating and the average student rating was calculated for each question and each class, and then a paired t-test was performed on the differences. The paired t-test is appropriate in this case (rather than the two sample t-test) since instructors and students within the same class are not independent. Although this method takes into account correlated student responses within an individual classroom, it still assumes that the average student responses from all of the classrooms are independent from each other. This may not always be the case; however, particularly where the same instructor may have taught several different courses or an instructor taught multiple sections the same course. Finally, this approach does not account for differences in the size of each class.

Clustered Data Analysis

A second method of analysis was also performed to account for correlated student responses within each classroom, between classrooms, as well as differences in class size. The instructor self-rating was compared to each individual student response using a random effects model. This clustered data analysis model employed a random intercept with the classroom as the cluster variable. This model does not require that the average student rating be calculated for each class. In this case, each student's rating is treated as a separate observation in the analysis. This analysis accounts for clustering within each classroom by assuming a common correlation within each classroom. The analysis was performed for each of the 16 items and their sum.

Discrepancy Analysis

Descriptive statistics on the differences between instructor and student ratings were performed using discrepancy analysis. Within each class, each student rating was subtracted from the corresponding instructor rating to calculate the numerical discrepancy between the instructor and the student. For example, if, for a given question, the instructor provides a rating of 3 and students #1 and #2 give a rating of 3 and 4, respectively, then the discrepancies for students #1 and #2 are $3 - 3 = 0$ and $3 - 4 = -1$, respectively. The percentage of students whose rating was X unit(s) different from the corresponding instructor rating was tabulated for each question. The possible numerical values for X (i.e., the discrepancy value) and the description for each are given below.

Discrepancy Value	Description
-4	The student rating is 4 points higher than the instructor rating
-3	The student rating is 3 points higher than the instructor rating
-2	The student rating is 2 points higher than the instructor rating
-1	The student rating is 1 point higher than the instructor rating
0	The student rating is no different than the instructor rating
1	The instructor rating is 1 point higher than the student rating
2	The instructor rating is 2 points higher than the student rating
3	The instructor rating is 3 points higher than the student rating
4	The instructor rating is 4 points higher than the student rating

The discrepancy analysis assumes that there are no correlations within or between classrooms (i.e., the same instructor did not teach several different courses or multiple sections of the same course).

To address the third hypothesis, summary statistics for the discrepancies were calculated to determine if the instructor and student evaluations differed based on the reason that the student took the course (e.g., major, distributional requirement or elective). In this case, the mean student and instructor rating for each question was calculated for each category (major, distribution requirement or elective). Differences in the mean student and instructor values were then calculated for each question (in each category) as well as the corresponding standard deviations and the ranges of discrepancy values.

Limitations of the Methodology

A limitation of the author's research method was not including any qualitative methodology techniques. Such a study could have examined the final section located at the bottom of the form that gives students the option to provide comments through words and prose.

The hypotheses and research questions in and of themselves were a limit of the study. Only a certain number of the items on this university-wide form were applicable to the study the author conducted. For example, item #10 states, "the instructor began and ended class on time." The timeliness of the instructor was not specified within the research issues. Since the structure and wording of the form was not tailored to the hypotheses set out in this research, the instrument itself was a limitation. The author needed to work within the confines of the form and essentially impose the study onto the provided items. This required that items not relevant to the hypotheses be disregarded when reviewing the study's results.

From a statistical perspective, a challenge was to examine a relatively large student sample with a significantly smaller instructor population. This required reducing the student population within the individual instructor's classrooms to a single number in order to draw comparisons. This led to a further limitation of the applied research methods: seven instructors did not elect to complete the self-evaluation form. This excluded, in turn, a relatively larger group of students from being included in the comparative data, despite the fact they completed the form

(n=150 students). This required running a separate set of statistical tests to assess the impact of the instructors' absent data amongst their classrooms.

When evaluating the third hypothesis—the reason students took a class—the number of students within each subset varied significantly: those who took it as a “Major Distribution Requirement” totaled 773, students who took the course to fulfill a “Distribution Requirement” totaled 109 and those who took the class as an “Elective” course totaled 36. Therefore, drawing conclusions will be distorted since the groups are highly uneven.

CHAPTER IV

THE RESULTS

Introduction

In the following section, the quantitative data were examined as it pertains to the research hypotheses. The chapter concludes with a review of the additional data analysis.

Descriptive Statistics/Data Analysis related to Hypothesis #1

The author hypothesizes that, using the same instrument, the instructors will have higher scores in their self-perceptions compared to their students.

Overall Instructor Versus Student Rating

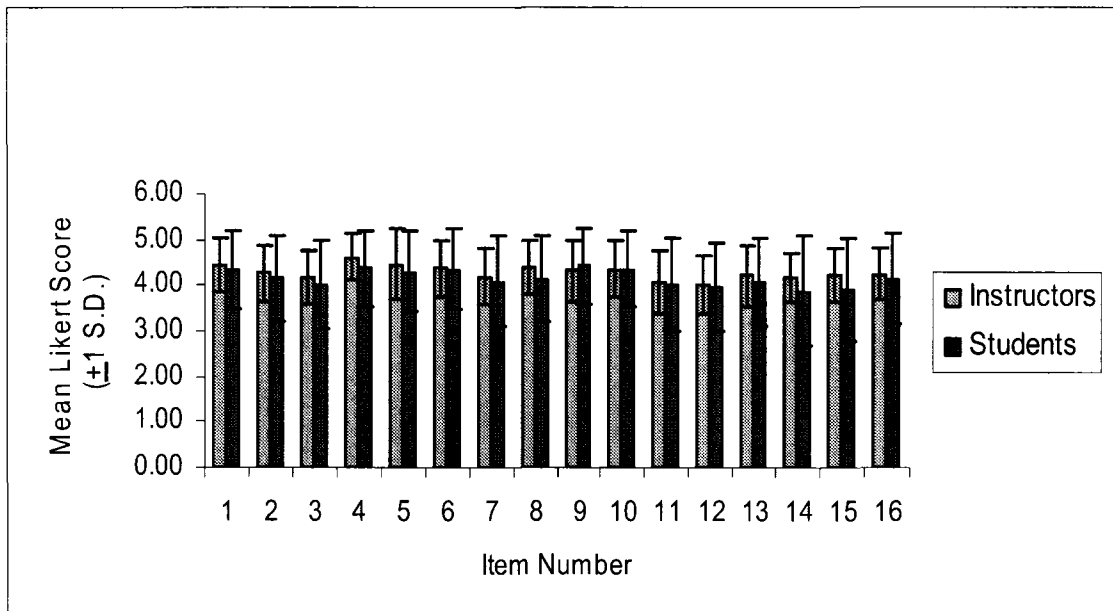
Table 1 (and Figure 1) shows the instructor and student mean Likert Scores (± 1 standard deviation; S.D.) for each of the 16 items on the teaching evaluation for all 79 classes. This analysis assumed that there was no correlation within or between classes and does not take into account that the same instructor may have taught several different courses or multiple sections of the same course. The p value reflects the degree of statistical similarity between the means of the two groups for each item. The mean values for the instructor scores are generally greater than the respective student scores, except for items 9 and 10, where they were lower. In spite of this trend, there were only five cases

Table 1. Instructor and Student Mean Scores (± 1 S.D.), Based on a 5-point Likert Scale, for the 16 Items on the Teaching Evaluations.

Item	Instructors		Students		<i>p</i> Value ¹
	MEAN	SD	MEAN	SD	
Item 1	4.46 (N=72) ²	0.60	4.36 (N=1447) ³	0.86	0.18
Item 2	4.28 (N=72)	0.61	4.17 (N=1442)	0.94	0.16
Item 3	4.18 (N=71)	0.59	4.02 (N=1440)	0.98	0.04*
Item 4	4.63 (N=72)	0.52	4.38 (N=1449)	0.84	0.0003*
Item 5	4.46 (N=72)	0.79	4.30 (N=1442)	0.88	0.10
Item 6	4.38 (N=72)	0.62	4.35 (N=1445)	0.89	0.73
Item 7	4.19 (N=72)	0.62	4.09 (N=1445)	0.99	0.17
Item 8	4.39 (N=72)	0.60	4.14 (N=1426)	0.93	0.001*
Item 9	4.32 (N=72)	0.69	4.42 (N=1433)	0.82	0.23
Item 10	4.35 (N=72)	0.61	4.36 (N=1445)	0.84	0.85
Item 11	4.07 (N=69)	0.71	4.01 (N=1240)	1.01	0.52
Item 12	4.03 (N=71)	0.65	3.96 (N=1445)	0.95	0.37
Item 13	4.22 (N=67)	0.67	4.09 (N=1442)	0.97	0.11
Item 14	4.17 (N=65)	0.55	3.86 (N=1423)	1.20	0.0001*
Item 15	4.23 (N=70)	0.57	3.93 (N=1439)	1.12	0.0001*
Item 16	4.26 (N=70)	0.58	4.15 (N=1440)	1.00	0.14

¹Student's t-test (2-tailed, with unequal variances); $p \leq 0.05$ (*) is considered statistically significant. ²Number of instructor responses is less than total number of courses taught (79) because not all instructors provided self-rating for all courses or all items. ³Number of student responses is less than total number of students enrolled (1,452) because not all students provided ratings for all items.

Figure 1: Instructor and Student Mean Scores (± 1 S.D.), Based on a 5-Point Likert Scale, for the 16 Items on the Teaching Evaluations. The graph is derived from the data given in Table 1.



(items 3, 4, 8, 14 and 15) where the instructors had statistically higher scores than those of the students. In no cases were the student scores statistically higher than those of the instructors. Thus, for the remaining 11 out of 16 items, the means of the instructor and student ratings were statistically similar.

Although the trend in the data generally supports the first hypothesis, it is only statistically substantiated for five of the 16 items.

Instructor Versus Average Student Rating

To account for the fact that student ratings within the same class are not independent, the average student response for each class was compared to that

of the respective course instructor *for each of the 16 questions* and the *sum of the 16 questions* via a paired t-test. The mean, median, standard deviation, range and p value of the differences between the instructor and average student rating was calculated for each item over all classes, and the results are shown below in Table 2.

Table 2. Instructor and Student Mean Scores and Mean, Median, Standard Deviation (S.D.), Range and *P* value of the Differences between Instructor and Average Student Likert Scores for the 16 Items on the Teaching Evaluations.

Item	Instructor	Student	Difference				
	Mean (N=72) ²	Mean (N=72) ³	Mean	Median	S.D.	Range	<i>p</i> Value ¹
1	4.46 (N=72) ²	4.40 (N=72) ³	0.06	0.15	0.64	-1.67 – 1.31	0.4011
2	4.28 (N=72)	4.21 (N=72)	0.07	0.03	0.69	-1.57 – 1.62	0.3945
3	4.18 (N=71)	4.07 (N=71)	0.12	0.07	0.62	-1.50 – 1.58	0.1009
4	4.63 (N=72)	4.42 (N=72)	0.20	0.25	0.58	-1.00 – 1.81	0.0038*
5	4.46 (N=72)	4.34 (N=72)	0.11	0.33	0.84	-2.64 – 1.63	0.2783
6	4.38 (N=72)	4.39 (N=72)	-0.01	0.13	0.67	-1.83 – 1.54	0.8613
7	4.19 (N=72)	4.10 (N=72)	0.10	0.03	0.69	-1.31 – 1.86	0.2047
8	4.39 (N=72)	4.17 (N=72)	0.25	0.34	0.62	-1.43 – 1.75	0.0011*
9	4.32 (N=72)	4.41 (N=72)	-0.10	-0.11	0.65	-1.67 – 1.45	0.1997
10	4.35 (N=72)	4.37 (N=72)	-0.02	-0.08	0.58	-1.50 – 1.04	0.7178
11	4.07 (N=69)	4.03 (N=69)	0.02	0.04	0.67	-1.67 – 1.42	0.7920
12	4.03 (N=71)	3.98 (N=71)	0.08	0.03	0.75	-1.40 – 1.64	0.3837
13	4.22 (N=67)	4.13 (N=67)	0.12	-0.04	0.72	-2.59 – 1.76	0.1790
14	4.17 (N=65)	3.92 (N=65)	0.29	0.24	0.64	-1.14 – 2.12	0.0005*
15	4.23 (N=70)	3.98 (N=70)	0.29	0.30	0.67	-1.29 – 2.09	0.0006*
16	4.26 (N=70)	4.20 (N=70)	0.08	0.00	0.67	-1.36 – 1.58	0.3384
Sum	67.33 ⁴	64.32	2.94	2.00	16.07	-45 - 60	0.1273

¹Student's paired t-test; $p \leq 0.05$ (*) is considered statistically significant. ²Number of instructor responses is less than total number of courses taught (79) because not all instructors provided self-rating for all courses or all items. ³Number of average student responses is less than that of the total number of courses taught (79) in order to coincide with the number of instructor responses in the paired t-test. ⁴Since the response to each item ranges from 1 to 5, the sum has a range of 16 to 80.

For example, for item 1, "*The instructor is well prepared for class*", on average the instructor rating was 4.46 while the mean of the average student response for the same classes was 4.40. The mean difference between the instructor and student rating was then 0.06 units. The median difference was 0.15 units, and the standard deviation of the differences was 0.64. For this item the difference between the instructor and student rating ranged from -1.67 to 1.31 (i.e., on one extreme, one class rated this item 1.67 Likert units higher than the instructor and, on the other extreme, one instructor rated this item 1.31 units higher than the average student rating for the class). The p value corresponding to the paired t -test for this item was 0.4011, which is not significant at the 0.05 level, indicating that there was no significant difference between the instructor and average student rating for question 1. It should be noted that student mean scores for each item in Table 2 differ slightly from those in Table 1. In this case, student scores were used for only those classes where the instructor also provided a rating.

Consistent with the results in Table 1, the mean values for the instructor scores are generally greater than the respective student scores, except for items 6, 9 and 10, where they were lower. At the 0.05 level of significance, there were significant differences in the instructor and average student ratings for items 4, 8, 14 and 15 (the boldface p values in Table 2), with the instructor rating significantly higher than the average student rating for these four items. Thus, for the remaining 12 out of 16 items, the means of the instructor and student ratings

were statistically similar. Although the trend in the data generally supports the first hypothesis, it is only statistically substantiated for four of the 16 items (the same items identified in Table 1).

Note that there was not a significant difference for item 3 in this case (as opposed to the corresponding result in Table 1), presumably due to the fact that a slightly different number of students were analyzed in each case. Furthermore, item 3 in Table 1 only bordered on significance in the first place. Although this method takes into account correlated student responses within individual classrooms, it still assumes that the average student responses from all of the classrooms are independent from each other. This may not always be the case, however, particularly where the same instructor may have taught several different courses or an instructor taught multiple sections the same course. Finally, this approach does not account for differences in the size of each class.

Clustered Data Analysis

To account for correlated student responses within each classroom and between classrooms, as well as differences in class size, a clustered data analysis was also used to test the first hypothesis. The results of this analysis are shown in Table 3. For example, for item 1, the instructor rating was 0.08 units higher than the student rating on average. The p value for this item was 0.4134, which was not significant at the 0.05 level; indicating that there was no statistically significant difference between the instructor and student rating for item 1.

Table 3. Results from Clustered Data Analysis Comparing Instructor and Student Likert Scores for the 16 Items on the Teaching Evaluation.

Item	Estimated Difference ¹	p Value ²
1	0.08	0.4134
2	0.09	0.4052
3	0.14	0.1921
4	0.22	0.0195*
5	0.13	0.2124
6	<0.01	0.9745
7	0.11	0.3208
8	0.26	0.0162*
9	-0.10	0.2900
10	-0.02	0.8707
11	0.04	0.7417
12	0.08	0.4439
13	0.12	0.2881
14	0.30	0.0295*
15	0.30	0.0171*
16	0.09	0.4061
Sum	1.49	0.2365

¹Estimated difference between instructor and student scores (i.e., instructor score – student score). ²P value from clustered data analysis; $p \leq 0.05$ (*) is considered statistically significant.

Consistent with the results in Table 2, the average values for the instructor scores are generally greater than the respective student scores, except for item 6, where it was approximately equal, and items 9 and 10, where they were lower. At the 0.05 level of significance, there were significant differences in the average instructor and student ratings for items 4, 8, 14 and 15 (the boldface *p* values in Table 3), with the average instructor rating significantly higher than that of the students for these four items. Thus, for the remaining 12 out of 16 items, the means of the average instructor and student ratings were statistically similar.

Consistent with the results in Table 2, the estimated differences between the instructor and student ratings in the clustered data analysis were comparable to those obtained by comparing the instructor and average student scores.

Although the clustered data analysis found statistically significant differences for the same items identified in Table 2, it should be noted that the p values in this case were less extreme than those obtained by comparing the instructor and average student scores (again as seen in Table 2). This is presumably due to the fact that taking into account the correlations within classrooms, between classrooms, as well as class size, results in an increased value of p (relative to the Student's t -test analysis of the data in Table 2), which is, although, still below the 0.05 level. Similar to the results shown in Table 2, the trend in the data from the clustered data analysis generally supports the first hypothesis; however, it is only statistically substantiated for the same four of the 16 items.

Discrepancy Analysis

Discrepancy analysis of the differences between instructor and student ratings is shown below in Table 4.

Table 4. Discrepancy Analysis of Difference between Instructor and Student Likert Scores for the 16 Items on the Teaching Evaluations.

Item	Difference Category ¹								
	-4	-3	-2	-1	0	1	2	3	4
	n %	n %	n %	n %	n %	n %	n %	n %	n %
Q1 (N = 1291) ²	0 0%	0 0%	25 1.9%	299 23.2%	582 45.1%	283 21.9%	67 5.2%	25 1.9%	10 0.8%
Q2 (N = 1288)	0 0%	0 0%	31 2.4%	310 24.1%	539 41.8%	268 20.8%	90 7.0%	42 3.3%	8 0.6%
Q3 (N = 1255)	0 0%	0 0%	25 2.0%	298 23.7%	503 40.1%	282 22.5%	108 8.6%	33 2.6%	6 0.5%
Q4 (N = 1293)	0 0%	0 0%	0 0%	230 17.8%	651 50.3%	305 23.6%	70 5.4%	23 1.8%	14 1.1%
Q5 (N = 1286)	0 0%	37 2.9 %	66 5.1%	237 18.4%	531 41.3%	303 23.6%	75 5.8%	31 2.4%	6 0.5%
Q6 (N = 1289)	0 0%	0 0%	42 3.3%	341 26.5%	546 42.4%	259 20.1%	61 4.7%	31 2.4%	9 0.7%
Q7 (N = 1289)	0 0%	0 0%	51 4.0%	350 27.2%	484 37.5%	260 20.2%	93 7.2%	31 2.4%	20 1.6%
Q8 (N = 1240)	0 0%	0 0%	25 2.0%	235 19.0%	550 44.4%	302 24.4%	84 6.8%	28 2.3%	16 1.3%
Q9 (N = 1277)	0 0%	2 0.2 %	54 4.2%	383 30.0%	567 44.4%	205 16.1%	42 3.3%	18 1.4%	6 0.5%
Q10 (N = 1289)	0 0%	0 0%	29 2.2%	309 24.0%	605 46.9%	265 20.6%	52 4.0%	21 1.6%	8 0.6%
Q11 (N = 1068)	0 0%	1 0.1 %	41 3.8%	275 25.7%	422 39.5%	234 21.9%	63 5.9%	28 2.6%	4 0.4%
Q12 (N = 1283)	0 0%	0 0%	70 5.5%	317 24.7%	486 37.9%	270 21.0%	94 7.3%	35 2.7%	11 0.9%
Q13 (N = 1235)	0 0%	14 1.1 %	35 2.8%	274 22.2%	500 40.5%	271 21.9%	92 7.4%	41 3.3%	8 0.6%
Q14 (N = 1173)	0 0%	0 0%	25 2.1%	277 23.6%	429 36.6%	245 20.9%	115 9.8%	68 5.8%	14 1.2%
Q15 (N = 1265)	0 0%	0 0%	26 2.1%	288 22.8%	460 36.4%	290 22.9%	113 8.9%	76 6.0%	12 0.9%
Q16 (N = 1267)	0 0%	0 0%	30 2.4%	329 26.0%	529 41.8%	232 18.3%	92 7.3%	42 3.3%	13 1.0%

¹Difference Categories: -4: Student rating is 4 points higher than instructor rating; -3: Student rating is 3 points higher than instructor rating; -2: Student rating is 2 points higher than instructor rating; -1: Student rating is 1 point higher than instructor rating; 0: Student rating is not different than instructor rating; 1: Instructor rating is 1 point higher than student rating; 2: Instructor rating is 2 points higher than student rating; 3: Instructor rating is 3 points higher than student rating; 4: Instructor rating is 4 points higher than student rating. ²Number of responses is less than total number of students enrolled (1,452) because not all instructors and students provided ratings for all items.

To assess the percentage of students that had instructor scores greater than or equal to those of the students, the appropriate categories in Table 4 can be collapsed into a single percentage. This is accomplished by summing n or % across the 0, 1, 2, 3 and 4 categories. For example, in item 1, there were 582 students (45.1%) whose ratings were equal to those of their instructors, 283 students (21.9%) whose instructor ratings were 1 point higher than those of the students, 67 students (5.2%) whose instructor ratings were 2 points higher than those of the students, 25 students (1.9%) whose instructor ratings were 3 points higher than those of the students and 10 students (0.8%) whose instructor ratings were 4 points higher than those of the students. Therefore, a total of $582 + 283 + 67 + 25 + 10 = 967$ students ($45.1\% + 21.9\% + 5.2\% + 1.9\% + 0.8\% = 74.9\%$) had instructor ratings that were greater than or equal to those of their students. For all 16 items, the percentage of students that had instructor scores that were greater than or equal to those of the students exceeded 65.5%.

Descriptive Statistics/Data Analysis related to Hypothesis #2.

Those instructors with the least discrepancies from their students' ratings will have higher overall student ratings compared to the overall student ratings of those instructors with more divergent scores.

The data in Table 2 were also used to test the second hypothesis. The 16 items listed in Table 2 were divided into two groups. The first group contains

those items where the mean instructor and student scores were statistically similar (i.e., items 1-3, 5-7, 9-13 and 16). The second group contains those items where the mean instructor and student scores were statistically different from each other (i.e., items 4, 8, 14 and 15). The mean and standard deviations of the student scores from each group were then compared for statistical differences. The results are shown in Table 5.

Table 5. Mean (\pm 1 S.D.) Instructor and Student Likert Scores for Statistically Similar and Dissimilar Groups of Items in Table 2.

Group	Instructors		Students		<i>p</i> Value ¹
	Mean	SD	Mean	SD	
1 (Items with statistically similar mean instructor and student scores – 1-3, 5-7, 9-13, and 16)	4.27 (N=12)	0.14	4.22 (N=12)	0.16	0.47
2 (Items with statistically different mean instructor and student scores – 4, 8, 14, and 15)	4.35 (N=4)	0.20	4.12 (N=4)	0.23	

¹Student's t-test (2-tailed, with unequal variances) was performed on the mean student scores from each group; $p \leq 0.05$ (*) is considered statistically significant.

Hypothesis 2 states that instructors with the least discrepancies from their student ratings (i.e., those in Group 1) will have higher overall student ratings than those where the instructor-student ratings are dissimilar (i.e., those in Group 2). Based on the results in Table 5, the mean student rating in Group 1 is indeed higher than that of Group 2. However, the difference is not statistically significant at the 0.05 level, and therefore, the second hypothesis is not supported by these data.

Descriptive Statistics/Data Analysis related to Hypothesis #3.

Those students taking the course as a requirement will be more critical of the professor than those students taking the course as an elective.

Summary statistics for the discrepancies in Table 4 are presented below in Table 6, based on the reason the student indicated for taking the course (i.e., major requirement, distributional requirement or elective).

Table 6. Descriptive Statistics on the Difference between Teacher and Student Rating Based on Reason for Taking the Course

Item	Reason	N	Mean	S.D.	Range
1	Major	688	0.11	1.01	(-2,4)
	Distribution	100	0.25	1.04	(-2,4)
	Elective	335	0.13	1.01	(-2,4)
2	Major	687	0.19	1.12	(-2,4)
	Distribution	101	0.32	1.11	(-2,3)
	Elective	334	0.14	1.02	(-2,3)
3	Major	680	0.16	1.12	(-2,4)
	Distribution	98	0.47	1.07	(-1,3)
	Elective	315	0.23	0.99	(-2,3)
4	Major	689	0.27	0.96	(-1,4)
	Distribution	101	0.23	1.00	(-1,3)
	Elective	335	0.25	0.89	(-1,4)
5	Major	686	0.06	1.21	(-3,4)
	Distribution	101	0.41	1.22	(-3,4)
	Elective	332	-0.02	1.11	(-3,4)
6	Major	688	-0.03	1.05	(-2,4)
	Distribution	101	0.29	1.26	(-2,4)
	Elective	332	0.14	0.98	(-2,4)
7	Major	687	0.08	1.15	(-2,4)
	Distribution	101	0.55	1.37	(-2,4)
	Elective	333	0.10	1.14	(-2,4)
8	Major	678	0.26	1.08	(-2,4)
	Distribution	96	0.48	1.15	(-1,4)
	Elective	308	0.22	0.98	(-2,4)
9	Major	687	-0.10	1.03	(-2,4)
	Distribution	101	-0.02	1.05	(-3,4)
	Elective	323	-0.15	0.89	(-3,4)
10	Major	687	0.07	0.95	(-2,4)
	Distribution	101	0.08	1.01	(-2,4)
	Elective	334	0.07	0.92	(-2,4)
11	Major	559	0.10	1.09	(-2,4)
	Distribution	84	0.07	1.05	(-2,3)
	Elective	283	0.05	1.05	(-3,3)
12	Major	684	0.03	1.13	(-2,4)
	Distribution	101	0.31	1.31	(-2,4)
	Elective	330	0.13	1.12	(-2,4)
13	Major	659	0.08	1.14	(-3,4)
	Distribution	98	0.39	1.34	(-2,4)
	Elective	322	0.30	1.11	(-2,4)
14	Major	629	0.35	1.21	(-2,4)
	Distribution	82	0.33	1.27	(-2,4)
	Elective	308	0.32	1.26	(-2,4)
15	Major	666	0.36	1.22	(-2,4)
	Distribution	100	0.54	1.32	(-2,4)
	Elective	333	0.26	1.15	(-2,4)
16	Major	676	0.11	1.10	(-2,4)
	Distribution	100	0.32	1.31	(-2,4)

	Elective	326	0.19	1.12	(-2,4)
--	----------	-----	------	------	--------

The number of students, mean, standard deviation and the range are presented for each question by reason. For example, for question 13, there are 659 students that were taking the course for their major, 98 students taking it for a distributional requirement and 322 students taking it as an elective. On average, the instructors rated themselves 0.08 units higher than the students that were taking the class for their major. For the students taking the class for a distributional requirement, the instructors rated themselves 0.39 units higher than the students. The instructors rated themselves 0.30 units higher than the students that were taking the class as an elective.

Of note, the standard deviations are presented in the column next to the means. The range for question 13 goes from -2 to 4. This reveals that at one extreme the student rating is 2 points higher than the teacher rating but at the other extreme, the teacher rating is 4 points higher than the student rating.

CHAPTER V

DISCUSSION

This chapter includes the following: an overview of the study, an overview of the research results related to the research hypotheses, a summary of the research results, the limitations of the study, the implications of the research, suggestions for future research and concludes with a summary section. Also included within some sections are observations and conclusions that are related to the literature on student evaluations of teachers.

Overview of the Study

During the last century, there has been a dramatic evolution in the usage of student evaluations of instruction in higher education settings. Wilson (1998) writes “only about 30 per cent of colleges and universities asked students to evaluate professors in 1973, but it is hard to find an institution that doesn’t today” (p. 2). What began as a voluntary and one-way process has grown to encompass a variety of new methodologies including the inclusion of teachers in the process. Over time, these instruments have been used as a means of measuring the quality of teaching. Most of the research has excluded the perspective of the instructor when examining the evaluation process. Relatively little research has included, and compared, the instructors’ ratings with their respective students.

This dissertation offers a model for expanding the application of teacher evaluations using both the student and the instructor perspectives. It is hoped that the results from this study will supply a new piece of quantitative data to the existing literature by comparing these two parties. As technology continues to advance, traditional teaching techniques may fall short of meeting the ever-changing needs of graduate students, especially those in the technical/computer science arenas. Educators and computer science experts may need to increase their involvement and collaborate on diverse and creative new teaching techniques. Such a partnership, potentially with the inclusion of students and instructors, could encourage the parties involved in the process to have meaningful input.

Attempts to bridge the gap between the expectations of the students and teachers, and the understanding that each party has a role and responsibility to communicate their needs, could create new strategies to enhance the relationship between the two parties. Viewing evaluations as an opportunity to advance change through two-way communication may be a valuable tool for growth that could help the evaluation tool make a more meaningful difference rather than just fulfilling an institutional requirement and/or chore.

To review, this study took place in a large, urban university setting in the adult part-time evening college within the computer science department. The total number of participants was 1,452. The instructors who participated in the study included adjunct and full-time professors and totaled 72. The classrooms that

participated totaled 79 courses. The university's "Course Evaluation Form" was used in this study. The evaluation was a "Likert Scale" standardized form consisting of 16 items. The forms were utilized for this study during the spring and fall semesters of 1996. The goal of the study was to compare instructor's ratings with their respective student ratings and draw conclusions to contribute knowledge to the research related to the course evaluation process.

Overview of the Results related to the Research Hypotheses

In this section, the quantitative results will be used to address the three research hypotheses. This will involve presenting each of the three hypotheses and following each with the significant related findings and previous literature that examined similar issues. The results of this study show an overall agreement of the students and faculty with a few exceptions where statistical differences were observed. Similar to Marsh et al. (1979), the author of this study found "that there was good agreement-both absolute and relative-between student evaluations and the corresponding evaluations by their instructors" (p. 157). Additionally, other researchers were consistent with the author's findings, such as Barnett et al. (2003), who write, "there were no overall differences between the scores for faculty members' self-evaluations and the scores for evaluations by the whole class of students" (Abstract section). The fact that they found "faculty member self-evaluations of their teaching and student evaluations of the same instruction produce similar results" is consistent with this author's research

results. They conclude, “faculty self-evaluations and evaluation ...of students can enhance the evaluation of faculty teaching” (Abstract section). Barnett et al. (2003) studied thirty-one faculty members over one semester. The research conclusions in this dissertation were also supported in part due to the large volume of participants in the study and the length of the study over two consecutive semesters.

The following research hypotheses were investigated in this study:

1. The author hypothesizes that, when using the same evaluative instrument, the instructors will have higher scores in their self-perceptions when compared to their students.
2. Those instructors with the least discrepancies from their students’ ratings will have higher overall student ratings when compared to the overall student ratings of those instructors with more divergent scores.
3. Those students taking the course as a requirement will be more critical of the professor than those students taking the course as an elective or a distribution requirement.

Hypothesis 1 Discussion

This hypothesis suggests that “when using the same evaluative instrument, the instructors will have higher scores in their self-perceptions when compared to their students.” On average, for all 16 items, the teacher responses and scoring was higher than the students but didn’t indicate a statistical difference. But when

taken individually, there were 5 items (3, 4, 8, 14 and 15) that did reveal statistical significance depending on the tests conducted on the data. Thus, it is important to consider the fact that these differences did not reveal themselves until further statistical and individual analyses were performed. This is consistent with the findings of Barnett et al. (2003) who write, “when differences at the level of individual instruction between ratings from faculty member self-evaluation and student evaluation of teaching occur, the small size of the difference, while statistically significant, may not be meaningful” (Discussion section, ¶1). Centra’s findings in his study (1973) support the author’s first hypothesis and results when he states, “the comparisons of the mean value [of students versus instructors] indicate that [the] instructors group generally rated or described their teaching more favorably than did their students” (p. 289). Centra (1973) further affirms the author’s findings regarding the first hypothesis when he concludes “there was also a tendency for teachers as a group to give themselves higher ratings than their students did” (p. 293).

The items revealing statistical differences were the following:

Item #3 - The instructor presents course materials clearly.

This item supports the author’s first hypothesis in that it shows some statistical difference between the students’ perceptions and instructors’ self-perceptions of the clarity in classroom presentations. The instructors in this study seem to perceive themselves as more lucid and intelligible in their lecturing skills than their respective students. Presenting material in an understandable

fashion seems central to the goals of an effective teacher. However, the course topic and content might be an extenuating issue in the hypothesis since this study specifically focuses on computer science related courses. Centra (1973) points out that instructors who teach in the area of the sciences “may feel that there is so much factual and theoretical material to cover in their courses that a fast pace coupled with a good deal of student effort is a necessity” (p. 294). He continues by writing “what teachers in the...sciences view as an acceptable pace and work load, however apparently does not coincide with their students, who frequently are using courses in other fields for comparison” (p. 294). This may especially be the case for students in advanced academic programs like the Master’s program that is the focus of the author’s study in this dissertation. The author speculates that when students are in advanced degree programs, the rigors and requirements of the courses are increased as compared to undergraduate programs. Also, for those students involved in this study, they are attending classes exclusively in the evenings and that may imply balancing other life roles as spouses, parents or full-time employees. The fact is that the Computer Science instructors involved in the author’s study may be unaware that their adult, part-time, evening students do not perceive the material as clearly understandable as the instructors intend. This represents a lack of connection in the student-instructor relationship that could be addressed once the problem is revealed.

Computer science is a growing area of education and is one of the reasons

the author selected this group in which to focus her study. Cimikowski and Cook (1996) write “technological changes are transforming society and the ways in which we learn” (p. 88) and in turn, the way in which our teachers must teach. Presenting materials clearly within this complex and changing field involves the ability to “demonstrate knowledge of uses of computers for problem solving, data collection, information management, communications, presentation, and decision making” (Cimikowski & Cook, p. 88) among many other skills. This combined with dealing with the unique needs of the adult learner that includes giving “students a more active, analytical role and [encouraging] them to take responsibility for their own learning” (Cimikowski & Cook, p. 90) makes for a big challenge for the instructor in the computer teaching field. Thus, the issue of presenting materials clearly becomes a challenge. The Center for Teaching and Learning (1997) suggests a protocol a teacher should follow in course presentations including setting clear course objectives, presenting the material “at an appropriate pace” (p. 3), developing the “students conceptual understanding” of the course topic and planning “assignments that solidify students’ understanding of the material” (p. 3). Once such a difference of perspective is discovered, like in the study reviewed in this paper, the protocols outlined can be incorporated. The awareness of a problem can be suggested as the first step in solving a problem.

Item #4 - The instructor is enthusiastic about teaching this course.

The results of this study revealed a statistical difference as related to

enthusiasm when comparing students and instructors perceptions. This item gives support to the first hypothesis that instructors will rate themselves higher than their corresponding students. To support this idea, it can be suggested that computer science teachers, untrained in educational techniques, might be weak in communicating a high level of enthusiasm when disseminating information. According to Watchel (1998), science and technical courses have instructors who “are less student oriented, the courses are less effective in presentation and faster paced, and the faculty are required to invest more time in research and seeking grants than their colleagues in other disciplines” (p. 197). The subject matter of computer science is likely going to affect the results of an “enthusiasm” rating in an inevitable fashion. Watchel (1998) writes “that a ‘poor’ teacher presenting interesting material is rated consistently higher on some dimensions of effective teaching than a ‘good’ teacher presenting boring material” (p. 197).

Research by a number of authors exists on the variable of instructors’ levels of enthusiasm that refute the author’s hypothesis. For example, enthusiasm levels of instructors were examined in a study conducted by Bosshardt and Watts (2001). Unlike this dissertation, they found that “instructors and students’ ratings on the enthusiasm item are fairly highly correlated” (p. 14). They add, “it seems likely that students are in a better position to judge an instructor’s ability to...teach with enthusiasm than they are to judge grading rigor or how well prepared an instructor is in a subject that most students have not seen before” (p. 14). Bosshardt and Watts add that “instructors...speaking ability and

enthusiasm are closely linked to self-ratings of teaching effectiveness...and students also value these traits” (p. 3). Another paper by Cimikowski and Cook (1998) address the issue of instructor enthusiasm as part of its content.

Cimikowski and Cook write, “since the instructors also believe in the relevance and importance of the material and this is reflected in their teaching, the students are continually commenting on the enthusiasm of the instructors” (p. 93).

Another study that positively correlates student and instructors with levels of enthusiasm of the instructor is described by Griffin (1998, Monitoring and Improving Instructional Practices section, ¶8) and conducted by Williams and Ceci (1997). These researchers “investigated whether changes in presentation style-increased enthusiasm-would lead to better student evaluations” (¶9). They found “that for every category of instruction rated, those students exposed to the more enthusiastic lecture rated [the professor’s] instruction and course higher” (¶9). Griffin writes “in a review of the effects of enthusiastic teaching research [it was found] that more enthusiastic instructors received higher student ratings” (¶9). Suffolk County Community College’s Office of Institutional Research (n.d.) writes that a variable that was “correlated with student ratings [and] enhanced learning [was] instructor enthusiasm and expressiveness” (Evaluation and Control of Potential Bias section, ¶2). Feldman concludes in his overview of a number of comparative studies related to student and teacher ratings that “they were similar in attributing high importance to the instructor’s enthusiasm” (p. 311).

In direct support of the author's first hypothesis, and in contrast to the previous research cited, Siegel and Johnstone (1985) gave attention to the issue of enthusiasm and how it relates to the computer science instructor specifically. They write that computer science instructors "frequently lack an appreciation for the dramatic aspects of the [teaching] craft, and for the important impact that flair and enthusiasm can have on learning" (p. 6). As a result of this lack in enthusiasm, computer science students are "sometimes subjected to poor presentations of subjects they thought were of great interest" (Siegel & Johnstone, p. 7). Consequently, this item (#8 on the form) of statistical difference revealed in the author's study may lead to the challenge of enhancing the student-teacher relationship by increased levels of enthusiasm on the part of the instructors.

Item #8 - The instructor's criteria for grading are fair.

In support of the author's hypothesis, statistical differences were found between students and their teachers when it came to the issue of grading fairness. As with the other items, instructors perceived themselves as being fairer with regards to grading than their students. Students appear to be sensitive to the issue of fairness in general in their education and especially as it relates to their grades. For example, Kaufman (1981) compared student ratings from a number of different academic areas and concludes, "the low level of 'Fairness' that art students indicated would be satisfactory for the ideal teacher indicates that fairness is a less important quality for an art teacher than for the

computer science...teacher" (p. 6). Schmelkin et al. (1997) write "our own analysis of students' reactions to the teaching environment...found that issues of fairness...for students were paramount" (p. 589). However, grading leniency has also been reviewed in the literature as it relates to evaluations. For example, The Ad-Hoc Committee on Student Evaluations of Ramapo College of New Jersey (2001) states "students rate those courses most highly that produce the best grades for the least work" Survey of the Student Evaluation Literature section, ¶6). This is confirmed by Brodie (1998) who writes "the professor assigning highest grades with least studying received highest evaluation, including paradoxically teaching the most intellectually challenging course" (p. 1). Watchel (1998) writes, "at this time the consensus is definitely that there is a moderate positive correlation between expected grade and student ratings (students expecting higher grades will give more favorable ratings)" (p. 202).

In the literature there are several theories related to grading and student ratings of instructors. One theory is that the easier grader receives the higher scores. The other theory is that the "most effective instructors cause students to work harder, learn more and earn better grades" and, in turn, the instructors receive higher rating scores (Watchel, p. 202). Then there is the idea that "pre-existing student characteristics such as prior subject interest affect...student ratings" (Watchel, p. 202) in a positive fashion.

It is generally argued the harder the instructor's grading criteria is for students, the lower the student ratings will be on the evaluations. Watchel (1998) writes,

“the more strict grading standards led students to rate the instructor lower even on components of instruction unrelated to fairness, such as humor ...and attitude toward students” (p. 202). Many variables seem to be unknown when the student rates the instructor according to his/her concept of “grading fairness” but the results appear to be the same. Brodie (1998) summarizes this end result when he states “grades cause students to change their evaluations of professors” (p. 3). He goes on to cite a study by Snyder and Clair (1976) that “found that students who were randomly assigned higher grades rated the professor higher than students who were assigned lower grades” (p. 3).

Kaufman (1981) addresses the issue of grading fairness as it pertains specifically to the computer science instructor evaluation process and directly related to the group studied by the author of this study. He writes, “psychology deals so extensively with topics of testing and bias, and since psychology and computer science are involved with statistical methods, students in those areas would be likely to expect more accurate and fair grading practices from their teachers” (Abstract).

There are studies that support the notion that the harder the course and grading criteria, the better rated the course on evaluations. For example, Marsh and Roche (2000) write “the most effective ways for teachers to get high SETs are to provide demanding and challenging materials, to facilitate student efforts to master the materials, and to encourage them to value their learning-in short, to be good teachers” (p. 226). Marsh and Roche feel it is essential to “debunk

popular myths that student evaluations of teaching (SETs) are substantially biased by low workload and grading leniency” (p. 202). Northwestern University (1999) writes “the highest marks often go to the most challenging courses” and that it is necessary to “understand that intellectually challenging courses graded with high standards will produce the best results” (Limitations of Student Ratings section, ¶2). Lawall (1998) examines the issue of expected course grades and the leniency in grading by writing that “this is a very controversial topic; however, the majority opinion sees no significant biasing affect” (SEEQ Research section, ¶21).

Grading fairness is a very important issue that can also motivate and promote learning from students. Palmer (1990) writes, “teachers can give students a chance to have their work evaluated several times before it must be finished...and grading then becomes more a tool of learning and growth than a final judgment on the final product” (p. 11).

Item #14 - I would like to take another course from this instructor.

For this particular item, it would be difficult for an instructor to answer in a self-evaluation format. This is obviously a limitation of the form itself.

Item #15 - I would recommend this course to other students.

This item has similar challenges to the previous item for the instructor to rate for him/herself.

In sum, three of the five items found to be statistically different between the

professors and the students relate to specific qualities possessed by the instructors. The items are the following: clear presentation of the materials, instructor enthusiasm and grading fairness. The fourth and fifth items revealing statistical difference related to whether or not the students would take another course from the same instructor (item 14) and whether or not the student would recommend taking another course from this instructor (item 15). As previously stated, these last two items are difficult to analyze since it is challenging for the instructor to respond and rate themselves on a self-evaluation form as to whether they would take another course from themselves or recommend themselves for future courses.

This first hypothesis was verified within the present study—that instructors will rate themselves higher in general when compared to their students. This is consistent with the findings of Centra (1973), who concluded “there was a tendency for teachers as a group to give themselves better ratings than their students did” (p. 293).

Hypothesis 2 Discussion

Hypothesis two states “those instructors with the least discrepancies from their students’ ratings will have higher overall student ratings when compared to the overall student ratings of those instructors with more divergent scores.” Instructors in the author’s study that had higher overall ratings were indeed closer to their student’s ratings than those instructors with lower overall scores.

The author calculated the mean values of the five professors' and students' responses showing statistical significance and compared them to the mean values of the twelve items where no statistical differences were observed (4.22 and 4.12 respectively). Overall, statistically these results only point to a trend consistent with the author's second hypothesis, although technically the hypothesis was not statistically proven at the 0.05 level.

The previous research regarding this hypothesis was found to be inconsistent with the findings of this dissertation. Moses (1986) found "both highly and poorly rated lecturers showed large discrepancies between their self perception and student perception" (p. 76) not supporting the author's second hypothesis. Moses writes "self-evaluations focus staff's attention on their own perception as teachers, and possible discrepancies between self and student evaluation may then motivate staff to change" (p. 86). She adds, "the analysis clearly showed that overall neither the superior teachers nor the satisfactory or less than satisfactory teachers shared the perceptions students had of them as teachers" (p. 79). Moses concludes that a "discrepancy between self and student ratings make it more likely that staff will act on the information received from the students" (p. 82). However, she also suggests that lack of discrepancy can be a positive situation that allows the instructor to feel a sense of affirmation by the students. Moses (1986) concludes, "self evaluation and student evaluation may match and show that nothing much needs to be changed [serving as] confirmation of what we are doing...and [increasing] our confidence" (p. 83).

Divergences, like the ones observed in the author's study, between students and instructors were addressed in the research by Centra (1973). Centra's study looked at the issue of differences between teacher self evaluations and students ratings and found a "discrepancy between an individual teacher [self] rating and the mean rating given by his class" (p. 294). Centra found, as this author did, that teacher self-ratings were "better" (p. 287) than their respective students ratings. He suggests "as an aid to instructional improvement, teacher self-ratings might in fact be used in conjunction with student feedback as a means of highlighting discrepancies for the individual instructor" (p. 294). In his study the discrepancies were "related to student-instructor interaction, course objectives, and the instructor's openness to other viewpoints" (p. 294).

For the 11 items on the evaluation form that did not show any statistical significance, it is still important to examine each for their meaning. These items in effect contradict the author's first and second hypotheses. However, it can be argued that those items that oppose the researcher's hypotheses (by not revealing statistical differences) might be as interesting as those with findings supporting the hypotheses. The following is a presentation, followed by a detailed review, of the 11 items where no statistically significant results were shown:

Item #1 - The instructor is well prepared for class.

The instructor being organized and equipped to lead the class appears to a major issue for both students and instructors when it comes to the classroom experience. Because the author of this study did not find statistical differences related to this item, her results are consistent with several studies that indicate instructor and student agreement on the importance of preparedness. Tang (1997) writes “the results of the present study shows that the instructor being “well prepared for each class’ [is one of] the most important predictors of overall teaching effectiveness” (p. 383). Schmelkin et al. (1997) observe that computer science instructors, specifically, and their students “both place high importance on teachers being prepared and organized” (p. 589) when teaching a course. Kaufman (1981) compared computer science students with students from other majors and writes, “on the dimension [of] preparation, Sociology majors rated the teacher much lower than Computer Science majors” (p. 5). Also consistent with the author’s findings, Feldman (1988) concludes “students and faculty were similar in placing high importance on teachers being prepared and organized, clear and understandable” (p. 311).

Item #2 - The instructor presents clear course objectives.

Contrary to the author’s results, Moses (1986) revealed a statistically different rating (at the 0.5 level or greater) in a study comparing instructors with students related to presenting clear course objectives: “while staff were convinced they provided clear objectives for each session...six of the eleven classes...did not think so” (p. 80). Centra’s (1973) findings showed that often “instructors and

students did not agree on the clarity of the “course objectives and what [was] taught” (p. 289). Tang (1997) writes, “professors need to be aware of the content of their course materials, the context of teaching the course materials, and the student who receive knowledge, skills, and information in the teaching process” (p. 384). Thus, the instructor needs to be organized, specific and clear when presenting the course objectives to his/her students. On the contrary, Feldman’s (1988) findings indicated, “faculty and students alike said that the clarity of course objectives and requirements was of low importance to good teaching or effective instruction” (p. 311).

Item #5 - The instructor encourages questions and class discussion.

Consistent with the author’s findings, the above statement did not show any difference between students and teachers in the study conducted by Moses (1986). Results indicated, “there was most agreement between staff and student perception concerning question 8, ‘There were enough opportunities to ask questions’” (p. 80). Feldman (1988) also concludes, in his overview of a number of comparative studies related to student and teacher ratings, “both groups [students and teachers] generally placed moderate importance on the teacher being open to class discussion and the opinions of others” (p. 311).

Like the author of this study, Kaufman (1981) addresses computer science educators, specifically, in his research related to item #5, regarding classroom questions and discussions. He writes, “Computer Science represents a relatively finite set of information, and assignments are given regularly so that the teacher

would have to be well prepared to answer specific questions related to the work assigned” (p. 7). Centra’s (1973) findings in his study, in contrast to the author’s findings, showed that instructors and students did not agree on “the extent to which students [were] free to ask questions or give opinions in class” (p. 289).

Item #6 - The instructor demonstrates mastery of the course materials.

Kaufman (1981) addresses the issue of subject knowledge in his study of student ratings from students representing different academic majors. He specifically looked at computer science majors as did the author of this study. Consistent with the author’s findings, Kaufman found that “for the dimension of Knowledge...sociology majors expected their teacher to be less knowledgeable than computer science instructors” (p. 6). When it relates to computer science students, it is clear that knowledge of the material is a high priority for both students and faculty. Also consistent with the findings of the author, Feldman (1988) concludes in his overview of a number of comparative studies related to student and teacher ratings that both groups “were similar in attributing high importance to...his or her knowledge of the subject matter” (p. 311).

Item #7 - The instructor’s criteria for grading are clear.

Many faculty, including computer science instructors, struggle with the issue of clarity in grading as it relates to their students. In the author’s results, grading criteria did not show a statistical difference worth noting. However, all of the research the author found contradicts the first hypothesis presented in the study and the author’s research results. The research indicates that clarity in grading

can often be a significant issue on the part of students and instructors. For example, Siegel and Johnstone (1985) write, “precise to an extreme about the substance of what they [computer science instructors] are teaching, these instructors are too often imprecise about their grading policies” (p. 6). They add that computer science instructors “fail to realize the importance the students place on knowing the various components of their grades, the relative weighting of the components, and the exact configuration of grading scales” (p. 6). Siegel and Johnstone (1985) suggest that students are frustrated “by unclear grading policies” (p. 7) and recommend a workshop that stresses “the importance of clear grading policies” (p. 8). In addition, Moses (1986) corroborates this idea from her comparative research findings when she writes, “students did not think that the lecturer had made [course] assessment requirements as clear as the lecturer had thought” (p. 81). Finally, Centra’s (1973) findings showed that instructors and students did not agree on “the extent to which instructors informed students of how they would be evaluated” (p. 289).

Item #9 - The instructor returned assignments and tests promptly.

The author of this study was unable to find literature or research studies related to this particular evaluation item.

Item #10 - The instructor began and ended class on time.

The author of this study did not find support for her hypothesis that instructors would rate themselves higher when it came to the issue of beginning and ending the class in a timely fashion when compared to their students. Furthermore, the

author of this study was unable to find literature or research studies that directly addressed this particular evaluation item. There was one study that looked at the factor of “class meeting time” and what, if any, relationship existed between that variable and student ratings. Watchel (1998) cites a study “by Koushki and Kuhn (1982) which found that very early morning classes, very late afternoon classes, and classes shortly after lunch receive the lowest ratings” (p. 196).

Item #11 - The instructor was sufficiently available for the consultation outside of class.

No statistical difference was observed in the results of the author’s study between students and instructors as related to instructor availability outside of the classroom. In contrast to the author’s results, a similar statement (“The lecturer seemed willing to offer individual help”) was shown as significantly different when comparing instructors and students in the research conducted by Moses (1986, p. 81). The results indicated, “while [the instructors] received good ratings on this item, their willingness was not equally apparent to all students” (p. 81). This willingness was even less apparent to students in a study conducted by Reid and Johnston (1999) who write, “teachers demonstrated no awareness of students’ perception of the importance of their approachability as part of good teaching” (Discussion section, ¶1). They continue “staff already feel that they are supremely approachable, and consequently show little intention to alter their priorities with respect to approachability” (Discussion section, ¶5). Finally, and also in contrast to the author’s findings, Feldman (1988) concludes in his

overview of a number of comparative studies related to student and teacher ratings that “students place more importance than do faculty on teacher’s availability and helpfulness to students” (p. 301).

Item #12 - The amount of work required for this course is appropriate.

The amount of work required by the instructor did not produce any statistical differences in the current study between students and instructors. The issue of workload amount is directly addressed by Northwestern University (1999) when they examined factors that influenced the student rating results. They write that “workload/difficulty” amount is “positively related to student ratings; that is, more difficult well-taught classes receive higher marks” (Limitations of Student Ratings section, ¶1). Similarly, Marsh and Roche (1997) also addressed the workload issue in their research. They write, “the Workload/Difficulty correlation was in the opposite direction than that predicted as a bias (SETs were higher-not lower-in more difficult classes; SETs were lower in ‘Mickey Mouse’ courses)” (p. 1191).

Of interest, Watchel (1998) reviewed several studies that examined the relationship between course workload and student ratings and found conflicting results from the ones just reviewed. He cited one study by Ryan et al. (1980) that “reported the introduction of mandatory student ratings at one US Midwestern university led faculty to reduce course workloads and make examinations easier” (p. 197).

Item #13 - This course challenged me intellectually.

The author of this study did not find a statistical difference between students and instructors related to the level of intellectual challenge presented in the individual courses reviewed. However, in contrast, a number of studies cited the debate that exists between students and faculty as related to intellectual challenge. For example, Siegel and Johnstone (1985) write “computer studies faculty receive lower ratings than do...all other faculty on the variable of ‘stimulating interest’” (p. 6). Furthermore, Feldman (1988) concludes in his overview of a number of comparative studies related to student and teacher ratings that “students place more importance than do faculty on teachers challenging students intellectually and encouraging their independent thought” (p. 301).

Some researchers have commented on the fact that students yearn to gain new knowledge when they take a course. Dooris (1997) supports this conclusion when he writes “that students want to be challenged [and] that they want to learn” (Penn State: Quality of Instruction Study section, ¶1). He adds “the single most stunning finding in all of the student data reported here is that the most powerful predictor of students’ overall evaluations of a course was the amount they felt they had learned in the course” (Penn State: quality of Instruction Study section, ¶1). Finally, Northwestern University (1999) observes that when it comes to the issue of “workload/difficulty,” teachers “understand that intellectually challenging courses...will produce the best [student rating] results” (Limitations of Student Ratings section, ¶2).

Item #16 - Overall, I rate this instructor a good teacher.

Marsh et al. (1979) address the overall ratings for courses in their comparative study of instructor and students. In correlation to the author's study results, they found "differences between student and faculty self-ratings were not statistically significant for either the 'Overall Course' or 'Overall Instructor' ratings" (p. 156).

Hypothesis 3 Discussion

Hypothesis three states that "those students taking the course as a requirement will be more critical of the professor than those students taking the course as a distribution requirement or an elective." If this hypothesis is to be supported by the data, the majors would be expected to have the most divergent ratings, relative to the instructors, as compared to the other two groups (i.e., those students taking the course as either a distribution requirement or an elective). However, based on this criterion, the third hypothesis is not supported by the data. For example, in item 13, the mean ratings for students taking the course as either a distribution requirement or an elective were .39 and .30 Likert-scale points, respectively, less than the mean value of the instructors. However, for the students taking the course as a major requirement, their ratings were only .08 Likert-scale points less than the mean value of the instructors.

This trend was similar for items 1, 3, 6, 7, 12 and 16, where the student ratings of the majors were more similar to those of the instructor as compared to

those students taking the course as a distribution requirement or an elective. It should be noted, however, that differences between these different groups are well within one standard deviation and thus not likely to be statistically significant. For the rest of the items, the differences in student ratings between the student groups and the instructors were fairly similar, well within one standard deviation. Consequently, the mean values of the student groups, relative to those of the instructors, do not appear to be significantly different from each other, which again is inconsistent with the third hypothesis.

This hypothesis was not supported by the data and results of the author's study. Initially the author assumed that students taking a computer science course as a requirement might have increased interest in the subject and more invested in their overall perceptions of their instructor and consequently, would be more critical of the instructor; whereas, those students taking the course as an elective might not have as high expectations and therefore might be less critical when rating the instructors on an evaluation form. This notion is similar to Wachtel's (1998) findings when he addresses the reason a student is taking a course as an issue in his research. He states "researchers have found that teachers of elective or non-required courses receive higher ratings than teachers of required courses" (p. 195). He hypothesizes, "this may be due to lower prior subject interest in required versus non-required courses" (p. 196). Lawall (1998) also addresses this issue, writing that "required courses are rated lower than electives" (SEEQ Research section, ¶18). Dooris (1997) confirms this trend, and

confirms the author's hypothesis, writing; "the bulk of the evidence suggests that students who are required to take a course may rate it more poorly than do students taking it as an elective" (Research on Student Rating Instruments, ¶8). Additionally, in support of the third hypothesis presented by the author, Coburn (1984) adds "most of the reported research seems to support the belief that students who are required to take a course rate it lower than students who elect to take the same course" (Faculty Concerns and Research Findings, ¶9).

On the contrary, and in support of the findings of the author's study, Lawall's (1998) research concludes that it "does not support" the notion that ratings differ depending on the whether the student is a major or non-major in the course (SEEQ Research section, ¶19). Dooris's (1997) study, consistent with the findings of the current study but inconsistent with the third hypothesis, found no significant relationship between the reason for taking a course and the ratings on course evaluations. Dooris (1997) writes, "it appears that whether students are majors or non-majors has no effect on their ratings of a particular course" (Research on Student Rating Instruments section, #9). However, "this question has not been as deeply researched as some others" (Research on Student Rating Instruments section, #9). Dooris notes in his study that "students taking courses as electives rated the instructors [only] slightly more favorably than did students taking courses as requirements" (SRTE and Other Research Findings section, ¶1). He concludes that the reason for taking a course does not influence the overall rating of an instructor (Dooris, 1997), which confirms the findings of

the author of this study.

It is important to note that on the evaluation form, and shown in the table 4 results, the “reason for taking a course” includes an item called, “Distribution Requirement for College to Graduate.” In contrast to a required course or an elective course, a “distribution requirement” gives students a list of classes from which they are allowed to select their preference. This “distribution” course involves some ambiguity since such courses are neither required nor completely voluntary on the part of the student. Therefore, with regards to the author’s study, distribution courses were disregarded since such a selection lacked relevance to the third hypothesis.

Summary of the Research Results

This study has shown that, in general, comparing instructors’ self-evaluations and student ratings of the same course produce similar results. When individual items were examined, statistically significant differences were observed. The findings of this study in response to the research hypotheses attempt to provide further insights into the student-teacher relationship. It appears clear, according to the results presented in this study as related to the first hypothesis, that the need exists to reduce the differences in responses between the two parties, especially related to grading fairness, clear presentation of the course material and instructor enthusiasm. It appears that instructors perceive themselves more positively in general than their students.

The second research hypothesis suggested instructors with the most “disconnect” from their students would likely be perceived less positively overall when compared to other instructors. Those instructors with the higher discrepancies from their students were in fact overall rated lower than those instructors with closer scores to their students but not at a statistically significant level. Examining discrepancies between self and student evaluation may serve to encourage and illicit changes on the part of the instructor that otherwise may have been ignored or overlooked. Finally, and contrary to the third hypothesis, it is not clear that the reason for taking a course affects student ratings on evaluations.

It seems evident that support systems need to be put into place in universities, in addition to the course evaluation process, to help promote more communication between the two parties.

Limitations of the Study

There are limitations that must be addressed regarding any study that is undertaken, and this study is no exception. Again, a limitation of this study is that it was exclusively quantitative by design. Further research with a focus on the “prose” that often are contained in the “comments” sections of the evaluation forms would lend additional data and instructor feedback. Marincovich (1998) writes that the “comments” sections of the evaluations are “probably the most underutilized of all the data on the forms” (p. 9). She continues, “very little has

been written on the topic of analyzing students' written comments" (p. 9).

Carefully scrutinizing the content of such sections of course evaluations would add the unique perspective of individual differences that are inherently limited in the "Likert" scale modality. It puts a unique "fingerprint" on each response intrinsically unavailable on a typical evaluation forms. Seldin (1993) advocates "several open-ended questions...be included on the rating form to allow student to respond in their own words" to "provide clues that clarify the underlying reason for particular ratings or that point to needed changes" (p. A40).

By nature, a quantitative piece of research, limits the affective and subjective interpretations of each of the variables. The focus on numerical results does not explore the motivations or "psychological agendas" of the students or instructors completing the surveys. For example, students may retaliate against "a faculty member who gave them harsh, blunt comments on their paper or homework" (Marincovich, 1998, p. 7) on an evaluation. A student might seize the opportunity to punish a teacher they simply do not like by giving the teacher low scores.

Because the form is subjective, some argue that teachers can encourage higher scores on their evaluations if, for example, their grading is lenient. Others argue that some of the most challenging instructors are given the most positive scores. The subjectivity of the responses is a limit of the study since evaluations can be viewed as formalized "popularity contests" (Tang, 1997). Coburn (1984) affirms this notion writing, "student ratings are measures of popularity rather than ability" (Faculty Concerns and Research Findings section, ¶1). Tang (1997)

argues that teaching evaluations, consequently, do not appear to reflect teaching effectiveness in an accurate manner. The results of a quantitative study of subjective material can risk misinterpretations and misuses of the results. Marsh and Roche (1997) recommend blending “qualitative research techniques...and quantitative techniques that have largely dominated SET research” to provide new sources of data. This idea is elaborated and taken further by Fries and McNinch (2003) who write, “the need for qualitative research-specifically, interviews and focus groups with students to better understand [student’s] attitudes towards SETS and their answers to specific items - continue to be timely and appropriate” (p. 342).

As reported in Chapter III, one very clear limit of this study, from a quantitative perspective, is that there is one subject (the instructor) being compared to relatively larger groups of subjects (classroom of students). This posed a statistical challenge in that the students as a group needed to be collapsed into a single number in order to conduct comparative analyses. This also suggests challenges in drawing conclusions that are accurate since the two numbers don’t reflect the same number of individuals. The author tried to account for this in the statistics that were chosen.

Another limitation of this study could be considered the evaluation form itself used in this study. England et al. (1996) write that typically items on evaluation forms “are not as applicable to the specific course as would be the case if the teacher developed or chose items specifically about aspects of the course”

(Conclusion section, ¶3). The evaluation tool studied in this dissertation was a standardized form used by the university at large and therefore is generalized across disciplines. Some argue that evaluation forms should be individualized for each and every department's needs. The University of Michigan's Center for Research on Learning and Teaching (2004) states, "a uniform system discriminates against some faculty, so a plan sensitive to individual variation should be developed" (¶4). Kaufman writes "interpretations of student ratings should include an understanding of these departmental differences, and emphasize the problem of comparing instructors across departments" (p. 7). Watchel (1998) confirms this idea, writing that, "we feel that it would be useful to conduct research on the effects of course characteristics on ratings in individual subject areas" (p. 198). As previously suggested, it would be of interest to design a form tailored to the unique characteristics of the computer science department, instructors and course characteristics.

There are those researchers who think the primary limit of instructor evaluations is the process itself. Gray and Bergmann, (2003) write, "the reliance on evaluations is bad for the health of relations between students and faculty" (¶14). They add that educational institutions should "move toward getting rid of this inaccurate, misleading, and shaming procedure" (Gray & Bergmann, ¶16). Further, it is suggested that students evaluating teachers is "like asking hospital patients to judge [the] medical care they've received" (Schwarz citing Greenwald and Gillmore, 1997, ¶7). The way the doctor, or instructor in this case,

communicates the material may be more important than the material itself. This also includes the fact that the patients or students in this case, have less knowledge on which to judge the other parties involved in the process.

In this ever-changing technological era in which we live, one cannot look at the future of instructor evaluations without considering that they are likely to become automated in their design, dissemination and collection processes. The system reviewed in this study was “paper and pencil” for the students and soon could be considered arcane as technology catches up with the evaluation system. Such a new system could be conducted on computer and automated by design for both students and instructors. Recker and Greenwood (n.d.) cite such a technique “utilizing the Web and HTML 2.0 with forms [that] provide a convenient, point-and-click interface for collecting on-line student responses” (User Perspective Section, ¶1). They add “as the number of computer labs, campus modems, and classrooms with network drops increase, a cross-platform, networked, client-server evaluation system becomes the most viable [evaluation] solution” (Recker & Greenwood, Organizational Perspective section, ¶1). They explain “data are automatically collated, processed and logged” which “improves efficiency and reduces paper-related costs” (Organizational Perspective section, ¶1 and 2). The Center for Teaching and Learning (1994) even suggests a high-tech alternative for the comments section in computerized student evaluations is “to set up an account on e-mail as a type of electronic suggestion box” (Using Student Feedback to Improve Teaching section, ¶4). Tang (1997) writes, “with

major changes [in the future], it is expected that the measurement of teaching effectiveness will be changed dramatically in the future” (p. 387).

There is also the limiting issue of self-selection in the participants who chose to participate in the instructor self-evaluation process. At the university studied, the choice of completing the evaluation form is voluntary for the students as was the choice of the instructors to complete the self-evaluation forms. This may skew the results since the author is not privy to the reasons why those instructors who did not participate chose not to. Furthermore, the program is a master’s degree program focusing on technology and thus, is narrow in scope; the subjects are limited to one department of study (i.e., computer science) in one educational institution, involving one level of students (i.e., graduate students). Computer science instructors and master’s students are a very specific, self-selected population and this may limit the way in which these results can be generalized for adult learners. Adult graduate students in technical areas are not required to seek out education at this level. It is often a personal choice for career advancement or for personal enrichment. The students at this level might also represent higher socioeconomic groups or be from a more educationally oriented-backgrounds than the average student.

This leads to another limitation inherit in this study. The demographics of the university and the region of the country may have an impact on the results. Different locations in the United States and different universities within those locations might alter the outcomes. Also, as previously stated, cross-cultural

studies have indicated that responses to evaluations may even vary depending on the country in which one resided.

Suffolk County Community College's Office of Institutional Research (n.d.) looks at two weaknesses related to student ratings; the "error of central tendency" and the "Halo Effect." Weaknesses due to "the error of central tendency" occurs because most people tend to avoid the extremes in rating, so ratings tend to accumulate in the center of the scale" (Limitations of Student Ratings, Students as Raters, and Application of Ratings Information section, ¶2). A common second problem, the "Halo Effect refers to the tendency of raters to be unduly influenced by a favorable or unfavorable general opinion of the person being rated and then...let that opinion color all specific ratings" (Office of Institutional Research, Limitations of Student Ratings, Students as Raters, and Application of Ratings Information section, ¶2).

The author of this study views one of the significant limitations of the study to be that the professors involved have not been required to have any formal training in the art of teaching and education. This problem is summarized by Stevens (1987) who writes,

"since few institutions provide explicit training on instructional methods as part of their graduate programs, we can reasonably assume that a large portion of college and university instructors, at some point in their careers, are deficient to some degree in the skills necessary for effective instruction or for effecting instructional improvement" (p. 35)

Making available, if not requiring, that the instructors in the computer science

field be educated to be effective communicators seems to be a large oversight. Such education could take the form of workshops related to presentation skills, syllabus construction, individualized learning styles and their impact on teaching and regular classroom visitations by trained educational professionals (Siegel & Johnstone, 1985). One of the pressing questions this author has is, what influence would such intervention have had on the evaluation data results?

The process of self-evaluation is a challenge, in and of itself, to ask any individual to undertake. The task assumes a certain level of self-reflection and honesty on the part of participant in the process. This assumption is a large “leap of faith” on the part of the researcher in this study and is perceived as an inherent limitation of the study as well. Moses (1986) addresses this issue when she writes, “self evaluation presupposes that we are able to look at what we do objectively and can assess the effect we have on students dispassionately” (p. 83). Another researcher writes, “there are factors inherent in self-evaluation data which make them somewhat suspect for dogmatic statements and possibly research” (Moses, 1986, p. 81).

Additionally, a limit of evaluations in general is they are often linked to job security rather than to skill enhancement, and this association produces pressures unlikely to motivate self-reflection. Some researchers suggest student evaluations are predominantly used to provide data regarding the “quality of faculty’s teaching to administrators who must make important decisions on granting of renewal, tenure or [promotions]” (Marincovich, 1998, p. 2). However,

fear and intimidation may be a more likely result. This idea is confirmed by Watchel (1998): “the use of student evaluations of teaching [can] reduce faculty morale and job satisfaction” (p. 193). In contrast, evaluations might be more useful for faculty if the instruments provided constructive and valuable feedback which would promote positive self-reflection and could result in new teaching techniques that would maximize their applicability.

In sum, there needs to be increased awareness of student evaluation “deficiencies, their limitations, and the circumstances under which they can be useful” (Ruskai, 1997, ¶6). Given the fact that these instruments are relied on in most universities and many of the forms are quantitative in their design, “at the very least...faculty should insist that any numerical component to the evaluation process used at their institution meet minimum standards of statistical validity” (Ruskai, ¶7). The classroom experience is the most important aspect of what is measured by evaluation forms. Feldman (1988) writes “what really needs to be known [about instructor and student evaluations] is how such similarities or dissimilarities come into play in the actual interaction between students and teachers in the classroom” (p. 314).

Implications of the Research - A Multi-Medium Approach

Because of the quantitative nature of this study, the results yielded statistical differences on a number of the items on the student evaluation form when compared with the instructor. These results generated questions about what the

findings mean as well as what the implications are for the future. The purpose of the study was to provide information regarding the way in which instructors perceive themselves versus their students. It was hoped and anticipated that the information would aid in the identification of differences in order to more fully understand how the student-teacher relationship can be improved. With these goals in mind, the implications of the results are presented.

The major and most direct application of the findings suggests that universities might benefit by reviewing the course evaluation process as a whole. For example, the results from this study could be useful in the development and refinement of evaluation instruments. The results could also raise some new ideas about the process of communicating the results to the instructors and increasing the overall involvement of the instructors in the process from the onset. There could be an instructor who represents other instructors and acts as a "liaison" with the administration communicating and acting on the behalf of all instructors. In turn, a student could act in a similar role.

If, as it was hypothesized in this study, instructors view themselves more positively than their students, it is essential to help close the gap with increased skill enhancement and communication among the parties. One way to do this is to have more formal supports in place and available to instructors after receiving the results of student evaluations. For example, focus groups and strategic learning sessions might provide encouragement and change. Rewards (e.g., certificates) and programs (e.g., mentoring sessions) may help to eliminate gaps

between instructors (Marincovich, 1998). Also, “teaching centers” might help utilize “student evaluations for teaching improvement” (Marincovich, p. 9) by “providing one-one-one teaching consultation services” (p. 9) to “help in interpreting and acting on their teaching evaluation results” (p. 9). Marincovich suggests teaching centers “provide teaching consultation services,” (p. 8)...“assistance in interpreting students’ written comments” (p. 9) and “produce materials that expose faculty to more of the research and thinking on student evaluations” (p. 9). It seems clear that “although student ratings are an important source of data for the evaluation of teaching merit, they should not be the only source” and “cannot carry the entire burden” (Scriven, 1995, p. 4). This is confirmed by Centra (1996) who writes, “the solicitation of evaluations from a wide range of sources can only increase the richness of the data available” (p. 55). Schmelkin et al. (1997) write “research has shown that the most effective use of instructional feedback leading to improvement in teaching effectiveness occurs when faculty are assisted by a professional teaching consultant in interpreting the feedback” (p. 589).

The author of this study became personally and professionally aware of the anxiety of instructors associated with receiving and dealing with the subsequent consequences, in some cases, of student evaluations results. There were not any support systems in place to assist in assimilating the data for each instructor regardless of the results (positive or negative). The author, and many other researchers, suggests that the ideal manner in which to evaluate instructors is to

incorporate this multi-medium approach. A simple written instrument does not adequately or fairly review a course and its teacher. Cashin (1988) states, “writers on faculty evaluation are almost universal in recommending the use of multiple sources of data” (Introduction section). Cashin continues “no single source of data-including student rating data-provides sufficient information to make a valid judgment about overall teaching effectiveness” (Introduction section). Seldin (1993) confirms this idea: “student ratings should never be the sole basis for evaluating teaching effectiveness” (p. A40).

Other sources of information, in addition to the student ratings, could encourage broader perspectives and make for a more comprehensive model of teaching evaluation. Suffolk County Community College’s Office of Institutional Research (n.d.) writes, “student ratings provide the most help when combined in a comprehensive program including a variety of evaluations tools and systematic faculty development” (Limitations of Student Ratings, Students as Raters, and Application of Ratings Information section, ¶5). Specific suggestions include “classroom assessment techniques, peer review and collaboration, and assessment of learning outcomes, to name just a few” (Marincovich, 1998, p. 12). The Center for Learning and Teaching (1997) suggests “by all accounts, the best way to use student forms to improve instruction is to consult with a colleague or teaching specialist regarding the meaning of the student data” (p. 1).

Since teaching includes “activities broader than classroom instruction” (Flinders Foundations of University Teaching, 2001, ¶1), other mediums of

assessment may provide a more balanced perspective of the classroom experience. Other aspects of teaching include student advising, curriculum development, supervision of teaching assistants, laboratory, etc. (Flinders Foundations of University Teaching). Some other ways in which an instructor could be evaluated include: he/she could evaluate him/herself on a regular basis, colleagues could conduct peer reviews, alumni letters and surveys could be distributed or focus groups could be led (Flinders Foundations of University Teaching, 2001). Stevens (1987) writes, "simply providing feedback is an insufficient tactic for behavioral change in a milieu as complex as the college or university teaching environment" (p. 37). Stevens continues by advocating "a system of institutional support, reward, and training for instructional improvement." Such a process might encourage "the instructor [to] learn how to design and implement alternative instructional procedures in response to feedback" (Stevens, p. 37). This, in turn, would ideally translate into "a coherent system of instructional resources [that] must be easily available to the instructor" (Stevens, p. 37). Stevens argues that without such mechanisms, the "instructor may be unable to gain the knowledge and support that is necessary to effect change" (p. 37). Such a program could involve instructors having "support groups" and mentoring each other on how to address the complex issues related to being an effective instructor. Such a group could provide the opportunity for instructors to share knowledge and interact with others performing in similar roles.

An area where the results of this research could be useful is in the delicate student and teacher relationship in the classroom, on the “front line.” This is when the teacher is in the minority and the students are the majority. Miscommunications will inevitably occur, but no instructor wants to wait until the course is complete to discover the problems. Even if the system requires the evaluation during the last class, if an instructor wants to make changes, it may be more constructive to create opportunities for open and honest dialogs with the students at other times throughout the semester, with regular mid-semester reviews. Students might benefit from other vehicles or venues of communication in addition to end-of-the-semester evaluations through which to voice their concerns. Such venues made available earlier in the semester might also allow for modifications to be introduced during their classroom experience. This could *promote an environment conducive for interaction and open communication* between the students and instructors. Having the evaluations always at the conclusion of the course might need to be reconsidered.

The implications of this study suggest that instructors could benefit from an increased understanding of the course evaluation results. Faculty may view the data in a variety of different ways depending on how the results will be used. For example, Schmelkin, et al. (1997) cited a study by Ory and Braskamp (1981) that “found that faculty ratings of the quality (e.g., credibility, usefulness, accuracy) of different types of student feedback depended on whether the feedback was for their own self-improvement or for promotion purposes” (p. 577). Ideally,

universities could offer services to help those instructors determine the best ways to analyze and use the results in constructive ways and to make changes that improve the student-instructor relationship.

Suggestions for Future Research

This study has generated several possibilities for additional research. As with any study, the need exists for the results to be replicated to lend it credence. Given the relatively small amount of research comparing student and instructor evaluations using the same instrument, follow-up studies are necessary to confirm and strengthen the generalizability of the results from the current study. For example, a follow-up study could be conducted on these same instructors now that a number of years have passed. And, although this study covered two consecutive semesters, a longitudinal study of student and instructor self-evaluations over a multi-year period might reveal patterns and insights not gleaned over one academic year. Needless to say, there are still many questions left unanswered and new ones created by any study, including this one.

There are a number of ways to enhance the findings of the present study. Increased focus is needed on the associations that have been uncovered by the research results in this study. For example, a future investigation could examine the reasons behind the differences in responses between the two parties especially related to grading fairness, clear presentation of the course material

and instructor enthusiasm. Additional research could also examine the possible causes that contribute to the disagreement in scores for those instructors whose students perceived their overall teaching to be less positive when compared to other instructors. Further research could delve in more detail regarding the role that taking a course as a major versus taking a course as an elective plays in the instructors' evaluation ratings.

A set of questions that can be answered only by additional research relates to the design of the evaluation form itself. What if the author designed her own form that contained more relevant items for both parties involved? Would that one design change have dramatically altered the results? Future research could examine the nature of the instruments employed in such studies. For example, the focus of such research could examine the items in terms of the phrasing and wording of statements. A comprehensive review of each individual item might yield possible changes or edits of the items to assist in providing new information regarding the teaching process. Watchel (1998) states, "many student evaluation instruments contain inappropriate items...items which were ambiguous, vague, subjectively stated, or did not correlate with classroom teaching behavior" (p. 194). Designing and implementing a new document based on the polling of instructors and students might yield unknown and valuable information for the students, instructors and administration.

The author's study compared student and instructors, but further studies could focus exclusively on the students in terms of their responsibilities in the

evaluation and learning process. What about the students' roles in their own learning? Dulz and Lyons (2000) write, "the instructor is assumed to be the expert and the primary source of knowledge" (Conclusion section, ¶3). Dulz and Lyons add "evaluation instruments are silent on the subject of student objectives, roles or responsibilities in learning" (Conclusion section, ¶3). Alternative studies could look at the students' responsibilities when it comes to their "side of the street" in the learning and evaluation process. Dulz and Lyons write, "students' central concern is personal relevancy, not instructor behavior" and "they don't find...evaluation instruments to be of any great value to them" (Conclusion section, ¶4). Students' involvement in the evaluation process might enhance their sense of "ownership" when being evaluated. Reid and Johnston (1999) address this issue, writing that instructors have the "desire to empower students to take more responsibility for their own learning" (Discussion section, ¶4) and they advocate a "student-participative approach to teaching and learning" (Conclusion section, ¶2). A study designed to examine students' input regarding the evaluation process might be of interest in the future. Marincovich (1998) writes "students might themselves take on the attitude of coach and concentrate on direct, constructive, and practical feedback" (p. 7) they can provide to their instructors.

Additional research could examine evaluations with both instructors and students within the technical/scientific fields of study at other institutions of learning or with different age groups. Cimikowski and Cook (1996) write there is

“blatantly an inadequate technological preparation for future teachers with the current accelerating technological advances that are occurring” (p. 94). Perhaps it might also be worthwhile to look at undergraduate programs or two-year programs, such as community colleges, to explore how different academic institutions/levels might influence the results. In addition, the academic institution where this study took place is large and urban. Future research could examine smaller, more rural academic settings to explore the effect of a different environment on the data results.

Further research could take a variety of other forms. For example, research of a comparative nature among a variety of academic departments of study would lend additional breadth and perspective to this sort of study. Would the findings of this study be duplicated in the entire university setting?

Another example of a study that continues the author’s research could explore entirely new procedures for evaluating and improving teaching. According to Ruskai (1997), “student evaluations need to be much more carefully investigated” (¶6). Ruskai adds, “their deficiencies, their limitations, and the circumstances under which they can be useful all need to be thoroughly documented” (¶6). It may of interest to provide a control group with a mid-term evaluation in addition to an end-of-term evaluation and then compare the results of the ratings with a classroom that only receives an end-of-term evaluation. Senior (1999) writes that there is some “evidence that mid-term evaluations are substantial improvements over end-of-term questionnaires” (Mid-Term Evaluations: Valuable

Alternatives section, ¶1).

A research study using an intervention technique could be designed that might expand the author's current research. Such a study could provide education courses to instructors in a research group focusing on teaching techniques and communication skills. The study could involve one group of instructors receiving education workshops and one group that does not and then could compare the impact of the intervention on the students' and instructors' self-ratings. In support of this concept, "faculty members could be offered courses or workshops on improving teaching effectiveness [and] receiving recognition on performance reviews for having taken such courses" (Student Evaluations: A Critical Review, n.d, Other Approaches section, ¶2).

Individual biases of any sort can skew the results of a "subjective" form such as a student evaluation. Continued research exploring the role of "biases" on the data could expand on the current body of research and could be of interest in the future. For example, if a particular instructor assigned a greater amount of homework than other instructors, did that bias the ratings of that instructor in the course evaluation? Such a bias could be studied in detail. England et al. (1996), for example, found that "student ratings of teaching are higher for courses that are rated as requiring more work or that are more difficult" (Personnel Decisions section, ¶3).

Cross-cultural concerns could be a focus of future research. An international student, or instructor (of which many existed in the author's study), may bring to

the classroom different expectations about the instructor-student relationship.

The cultural background of an individual may influence the classroom experience. Different cultures may perceive the nature and purpose of education in vastly contrasting ways. Obviously, classrooms are comprised of heterogeneous groups of individuals, and this fact is going to be observed even in the course evaluation results. It may be useful to determine to what degree such a variable influences, or changes, the ratings on the student or, for that matter, on the instructor self-evaluation forms.

The study presented in this paper examined both full-time and part-time instructors. It may be of interest to see if any differences in student ratings emerge when comparing the two groups in future studies. Watchel (1998) cites a study that “found that adjunct faculty tended to give higher grades and receive higher [student] ratings than full-time faculty, even though most students were not aware of the of their instructor’s status” (p. 202).

A recommendation to the department in which the author worked, is that, in the future, the chairperson might benefit by regularly requiring the instructors to conduct self-evaluations. Thus, in addition to the student responses, the administration could view the instructor responses. Also, Olp et al. (1991) suggest, instructor self-evaluations “[provide] division chairmen with a printout on which the student mean scores and the faculty self scores for each statement are reflected” (p. 6). Olp et al. suggest that such data “allows for a dialog to occur” (p. 6) and provides additional information not currently available to departments

and department chairs.

Studies that are of a comparative nature should continue in a similar way to the study presented in this paper. Barnett et al. (2003) recommend, “for now, a more comprehensive evaluation process, incorporating both student evaluations and faculty member self-evaluations should be employed” (Discussion section, ¶2). The author of this study hopes that instructors can be encouraged to take a more active role in the evaluation process with the goal of learning and improving their teaching skills. This concept would be consistent with Barnett et al.’s research, which promotes the idea that “with regular input through self-evaluation, faculty may move from feeling noncommittal to favorable toward evaluation by students...increasing instructional improvements and ultimately [enhancing] learning” (Discussion section, ¶2).

The opportunities are extensive for additional research into the comparative studies of student and instructor self-evaluations and the evaluation process in general. The suggestions presented are intended to expand on the findings of the current study and offer options for new areas of exploration.

Summary

It is the hope of the author that the results from the current study have furthered the understanding of the factors that influence the teacher-student interactions. This study was quantitative in design and intended to explore the student-teacher relationship. Data were gathered through a university evaluation

program. The results suggested statistical differences related to a number of the items that may lend insight into the divergent areas that exist between students and instructors. The findings may help to indicate the usefulness and applicability of the evaluation system while also raising concerns about the weaknesses of relying on them too heavily. It appears that evaluation programs in higher education institutions in general have been, and will continue to be, utilized. Identifying the strengths along with the weaknesses of these evaluation programs might help to provide strategies that would reduce the risks of misinterpretation or the missed opportunity for valuable and constructive feedback.

What clearly becomes a challenge is how to design and implement an evaluation system that promotes and maximizes the student-teacher relationship and minimizes the discrepancies that can appear in comparative evaluation results. Creating options or new ways to examine the data related to evaluations will encourage institutions to “look outside of the box.” In turn, this could help students and teachers think and interact in new ways in the future. In order to promote these new strategies, students and teachers may need to be given the opportunity to become more involved in the evaluation process and take more “ownership.”

These research findings may have implications for the development of new evaluation forms that restate the items for greater relevancy for computer science departments. In addition, increasing the level of involvement of both parties

might enhance the overall system. The act of teaching technical material in a stimulating fashion and effectively communicating new information is not an intrinsic skill; it needs to be learned. Computer scientists might benefit from educational courses to help build those skills. Palmer (1990) writes “good teaching requires...the courage to expose one’s ignorance as well as insight, to invite contradiction as well as consent, to yield some control in order to empower the group, to evoke other people’s lives as well as review one’s own” (p. 16). Palmer also writes “good teachers dwell in the mystery of good teaching until it dwells in them...and as they explore it along and with others, the insight and energy of the mystery beings to inform and animate their work” (p. 11). University leaders may need to support efforts to actually evaluate the evaluation system and examine its relevancy in the current academic environment. The purpose and mission of the evaluation system on a whole, as well as the more minute details like the phrasing of each item, might also be reevaluated for their relevancy.

One of the main recommendations based on the results of this study is that there needs to be particular attention to both the students and the instructors in the evaluation process. This is not a process that can isolate either party. Both are involved and both have needs and vital perspectives that require attention. Leamon et al. (1999) writes “faculty and students differ significantly in their expertise, perspectives, and content knowledge of the subject being taught” (p. S24). This study was undertaken because of a perceived “disconnect” between

instructors and students when it came to the teacher-student relationship. There is considerably less research on comparative studies of student evaluations than those on evaluations in general. The author hopes that studies such as this one will expand the literature on this subject and broaden perspectives so that the instructor is included more in the evaluation process. Attention should be paid to how the results of evaluations are communicated to the instructor and how he/she can make constructive changes without being too disillusioned if the results are considered sub-standard.

This study ultimately hopes to make an important contribution towards identifying areas of divergence and statistical differences in the student-teacher relationship that may close some gaps. Building awareness of the components that are missing in the student-teacher relationship is just as important as affirming what is positive and constructive in the relationship. Obviously, in any relationship, there is always room for improvement, and part of the growth process is being amenable to making the changes necessary to rectify the situation. The desire on the part of both parties involved must exist to improve the process of evaluation and make it as applicable as possible. The intention of this study was, at minimum, to provide recommendations to instructors striving towards excellence in their teaching quality as well as in their interactions with their students.

One can argue that comparing student and instructor perceptions is like comparing “apples and oranges.” The author of this study has attempted to

show that combining both perspectives adds valuable insights into the traditionally limited process. Bosshardt and Watts (2001) confirm this idea: “the primary value of such comparisons may be to see whether instructors and their students perceive the same strengths and weaknesses in instructors’ teaching...[and if] students and instructors generally agree or disagree on the overall rating of teaching effectiveness” (p. 5).

This study argues that the more parties that are involved in the complex task of reviewing an instructor’s effectiveness, the more valid the data. Students’ ratings should be seen as one component of a multi-dimensional approach to the evaluation process. Ultimately, there is not any “‘silver bullet’ with respect to evaluation of teaching effectiveness” (White, 1995, p. 84).

The single strongest recommendation of the author is to provide those who instruct in the sciences, specifically in computer science, with education classes to enrich the instructors’ teaching skills. Siegel and Johnstone (1985) write that regarding computer instructors, “while enthusiasm and expertise are great assets to higher education, these...instructors are frequently novices in the classroom” and need “opportunities for the development of teaching skills” (p. 4). Computer Science is a difficult subject to teach and “faculty are often not creative presenters or communicators” (Siegel & Johnstone, p. 6). Since “they are trained in a highly technical fashion, these faculty may have a difficult time communicating with people in an engaging, well-paced and systematic fashion” (p. 6). However, the students in classrooms, no matter what their ages,

represent the future of our society. Teachers are vital for the transmission of information to these students, and this includes students in the expanding and proliferating field of computer science. The hope in the future is that the most productive means of evaluating instructors, and their teaching effectiveness, will be more clearly understood and that this study will contribute to that body of knowledge.

APPENDIX B

Informed Consent Form

Informed Consent Form

Instructor self-perception versus student perception:
A comparative study of computer science instructors in an urban
adult education program

I am presently a Doctoral student at Boston University in the Developmental Studies Department in the School of Education. My dissertation will attempt to compare technical instructors' views of themselves, versus their students, to determine variables that contribute to effective teaching techniques. This data will be statistical in nature and extracted from the Boston University official "course evaluation" forms. Initial indicators suggest there is a dearth of statistics on the relationship between computer science instructors and their students. Additionally, I plan to conduct a qualitative analysis of the inner thoughts, concerns and reflections of the same students and instructors using the "comments" section located on the back side of the evaluation forms. I intend to conclude with suggestions for future areas for related research and with ideas for facilitating positive relationships between instructors and their respective students.

Your participation will greatly assist my collection of information. As a participant, you will be asked to complete a course evaluation simultaneously to your students at the location and time allocated by the college. I am aware that time is a precious commodity for computer science professionals who teach at night, however, your involvement will make a valuable difference in my research. The names of all subjects will be removed, and replaced with an identification number, to protect and respect your privacy. You may refuse to answer any particular questions or withdraw your participation at any time. At the completion of my dissertation, I would be happy to send you an abstract of my conclusions.

For more information, please call 617 (723-2122).

_____ I agree to take part in this project. I understand what will be required of me and that I can withdraw my participation at any time.

Signature _____
Date _____

If you wish a summary of the research results, please include your address:

APPENDIX C

Approval Memo from College Dean

Boston University

Metropolitan College
Computer Science
755 Commonwealth Avenue
Boston, Massachusetts 02215
617/353-2566

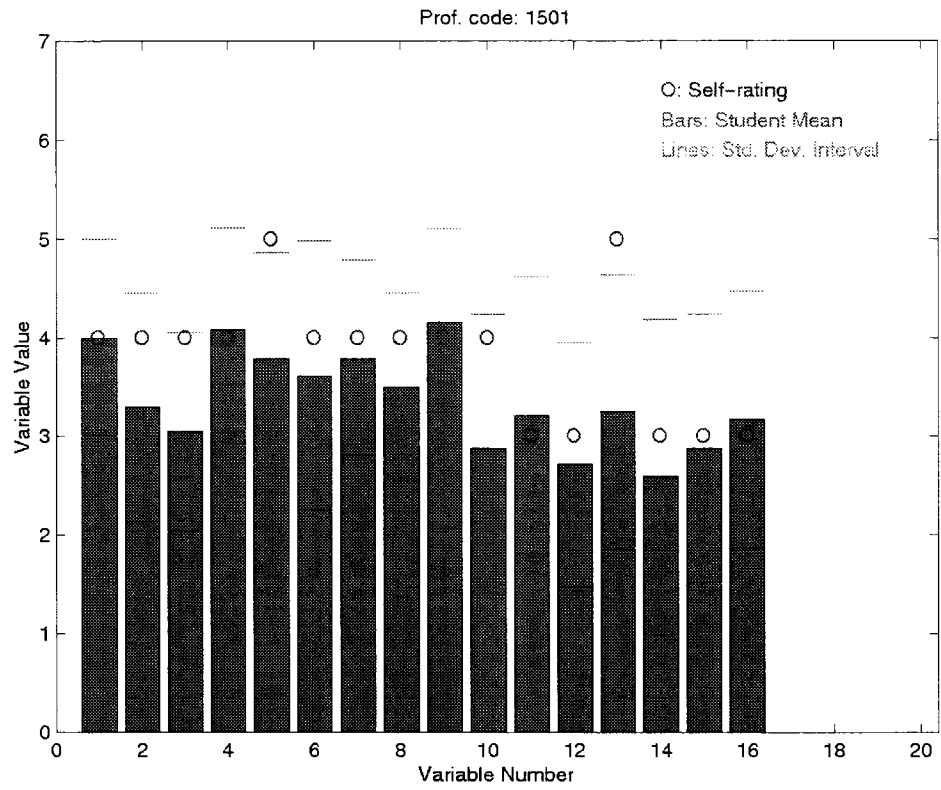


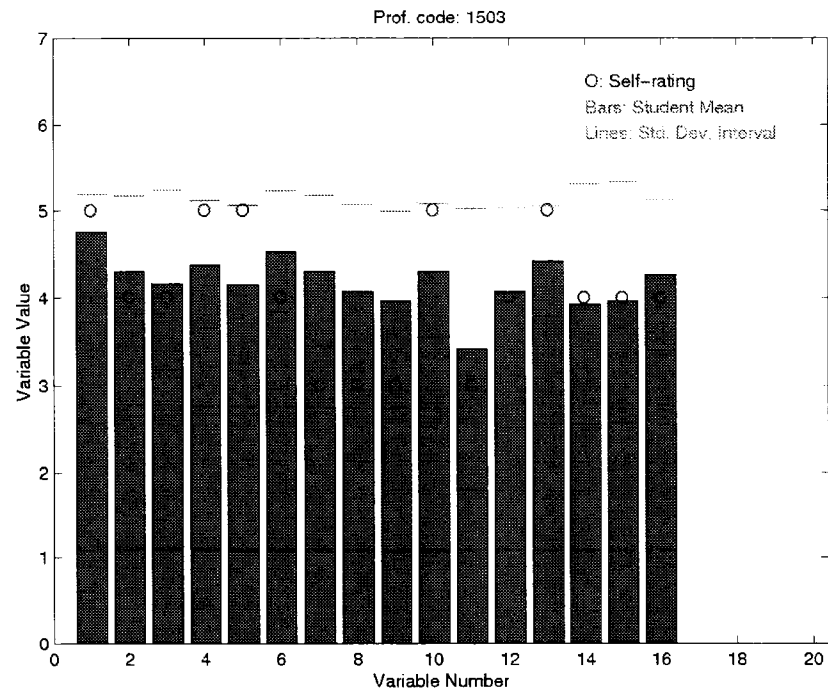
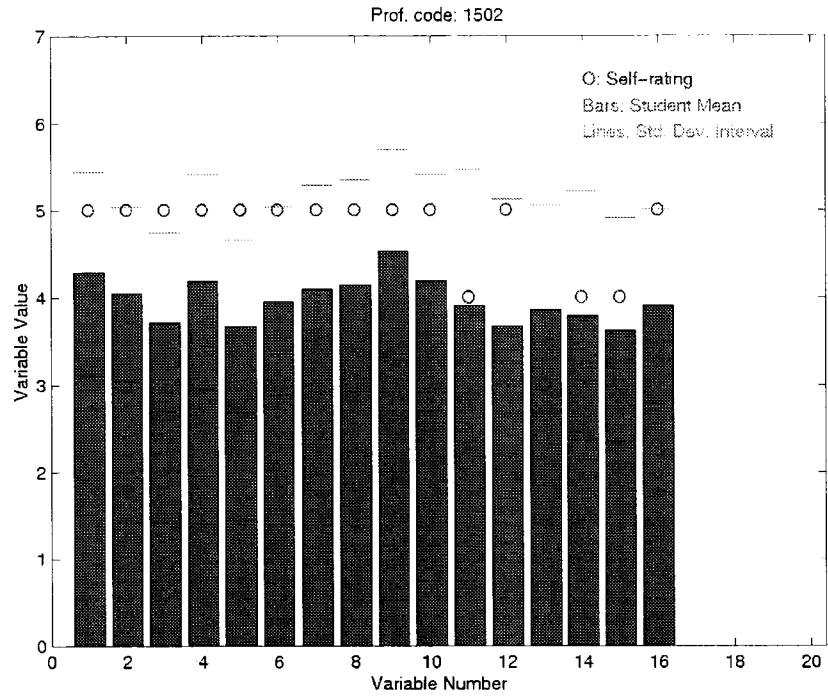
To: CS/CIS Faculty
From: Rom Skvarcius
Date: March 29, 1996

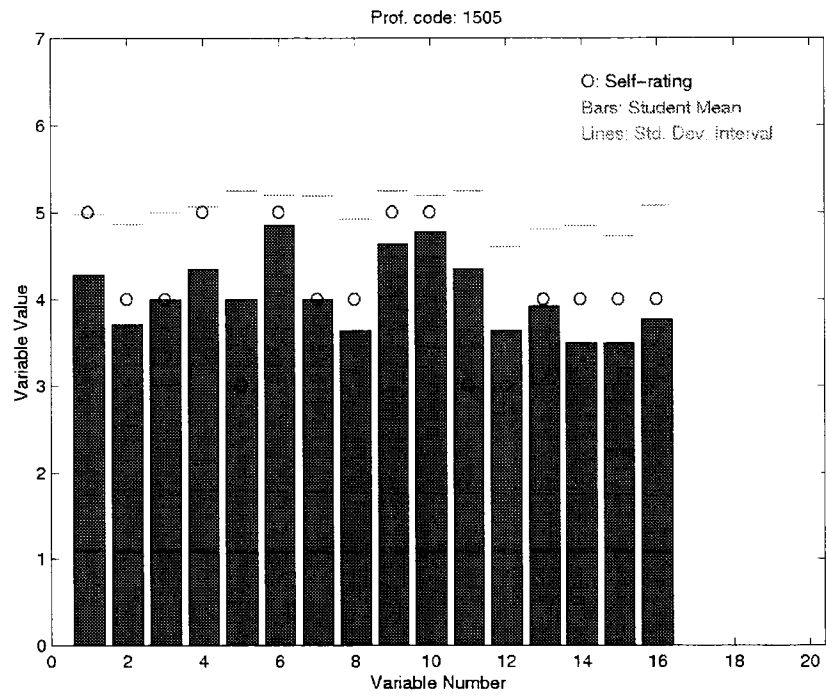
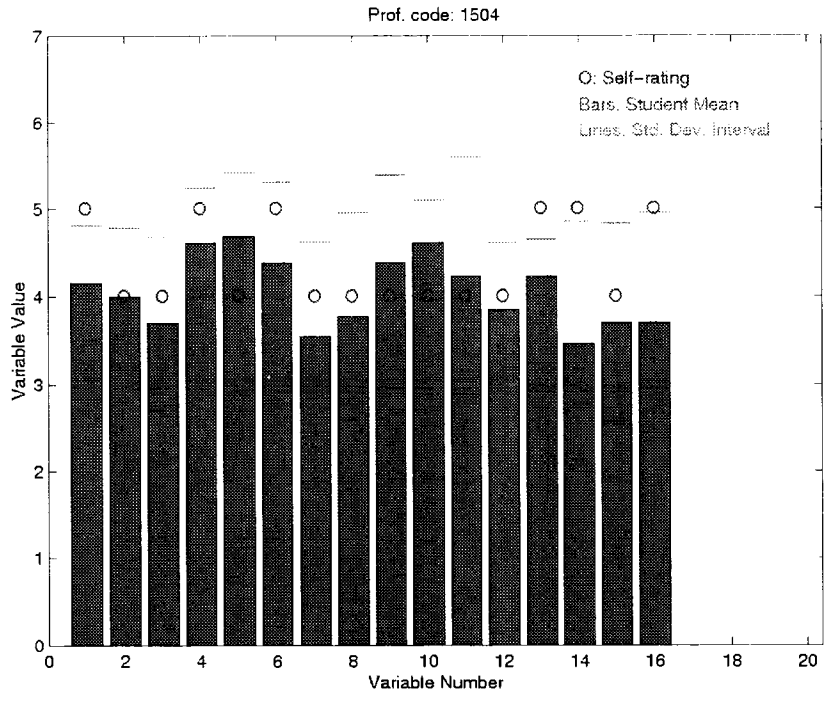
A handwritten signature in black ink, appearing to read "Rom Skvarcius", written over the "From:" line of the memo.

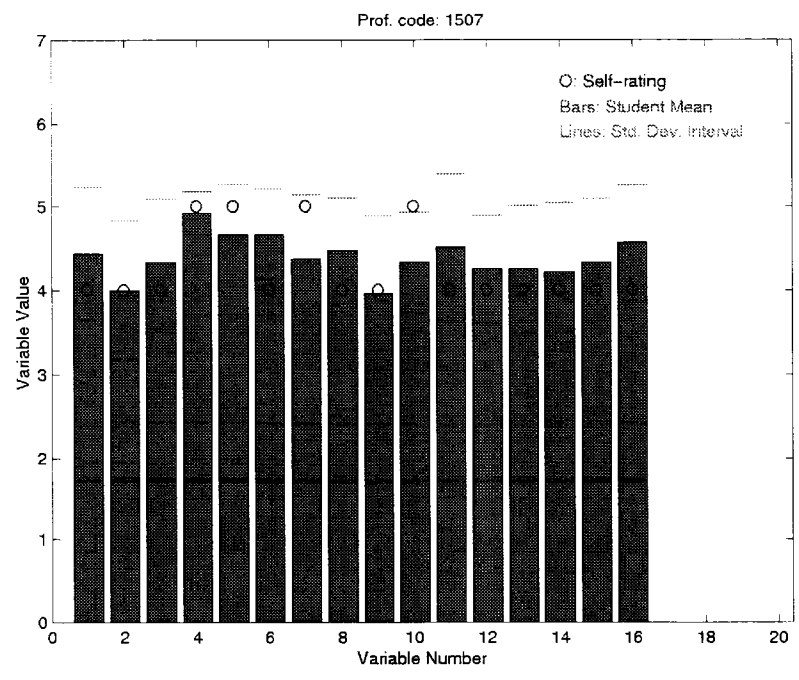
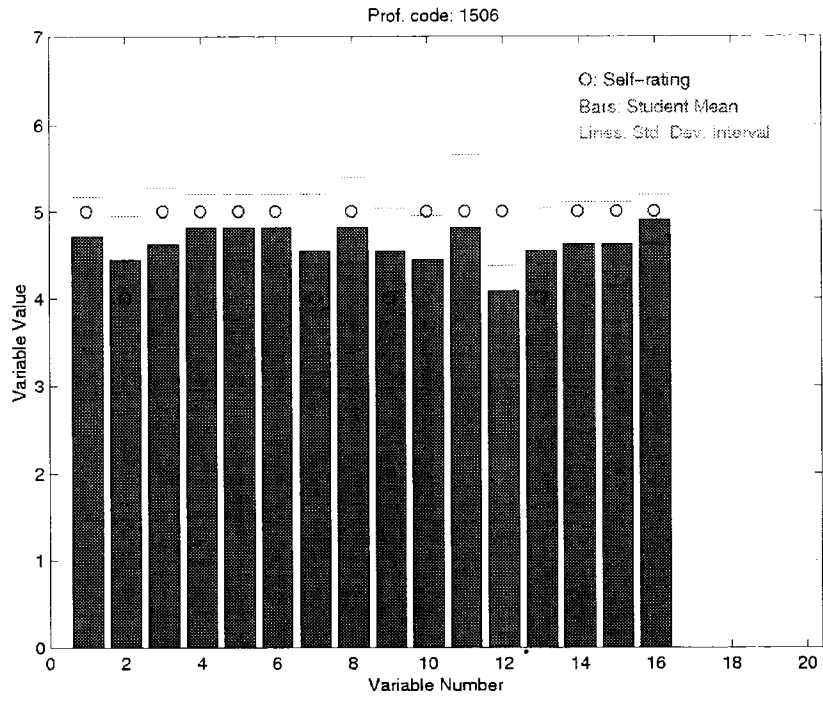
As many of you are already aware, Laurie Schwartz is in the process of completing her doctoral studies in education. Her dissertation will explore instructor self-perception when compared to students. Would you please fill out your own evaluation form to assist her in conducting her research. This material will be treated with complete confidentiality.

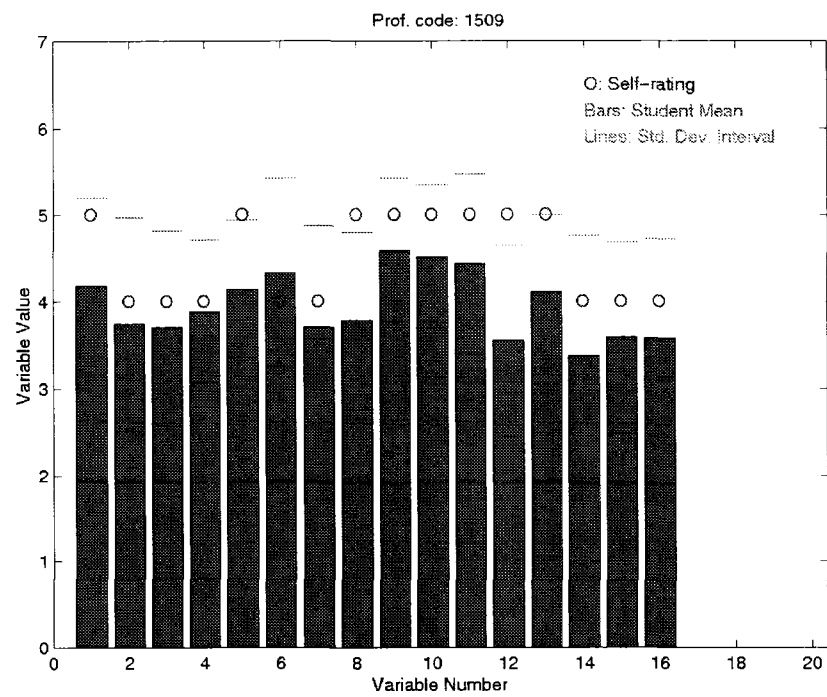
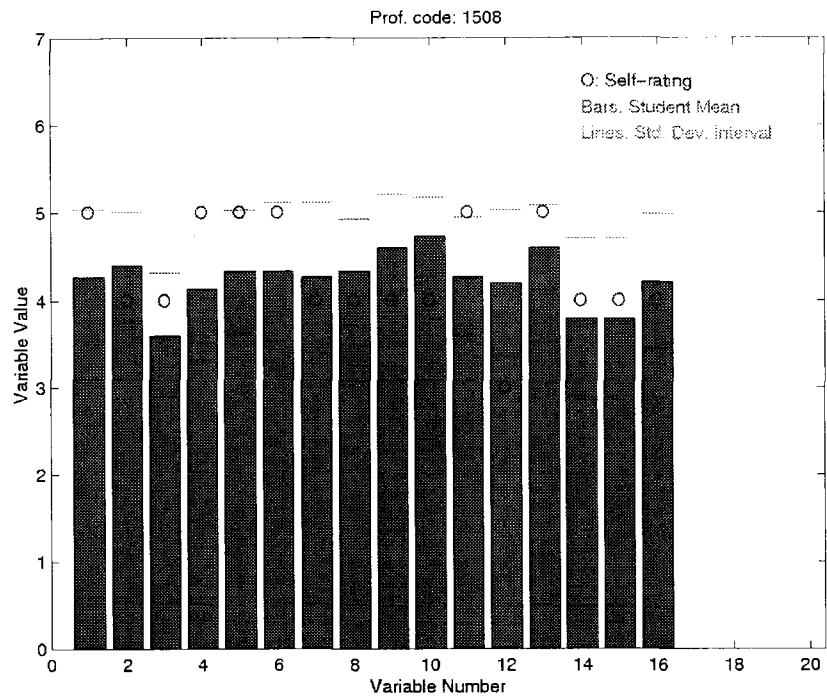
GRAPHS

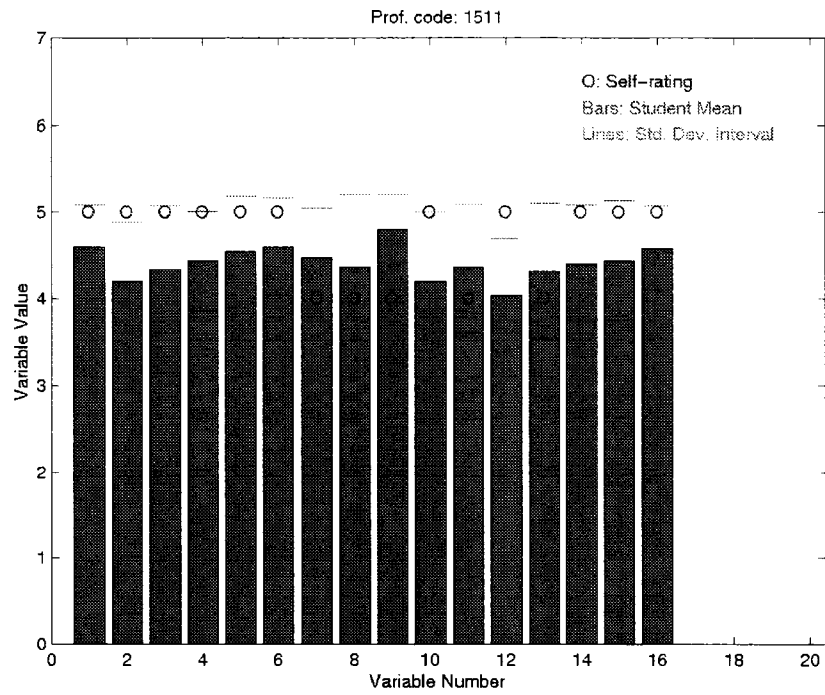
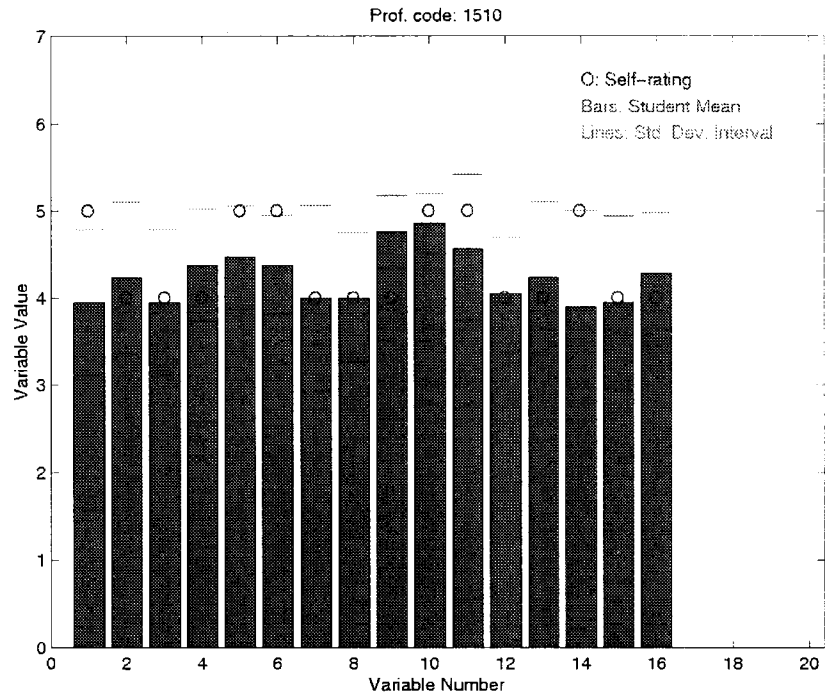




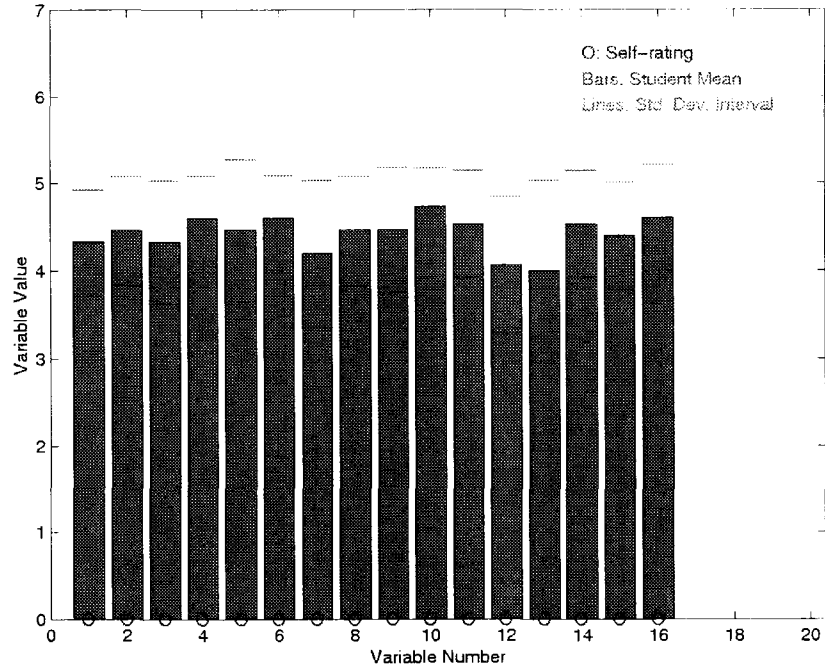




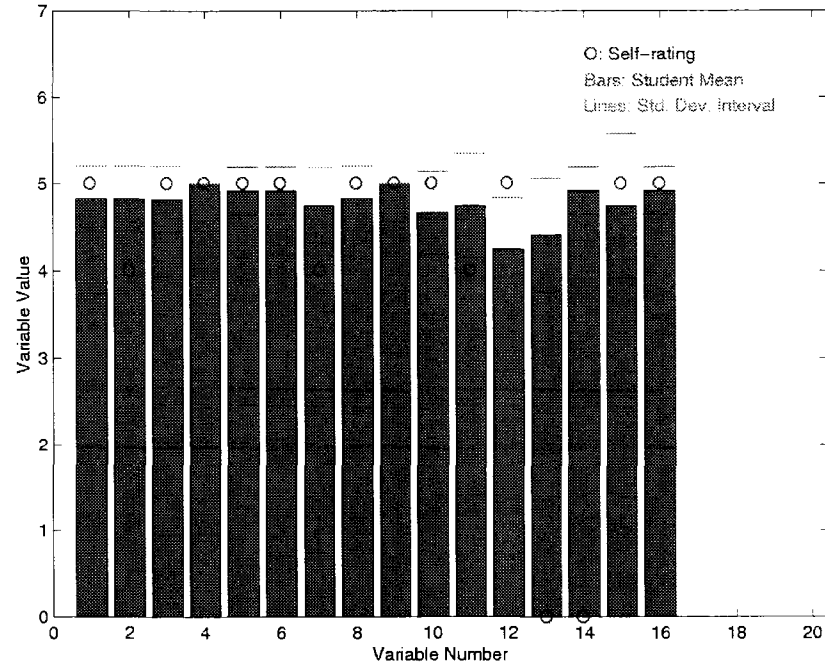


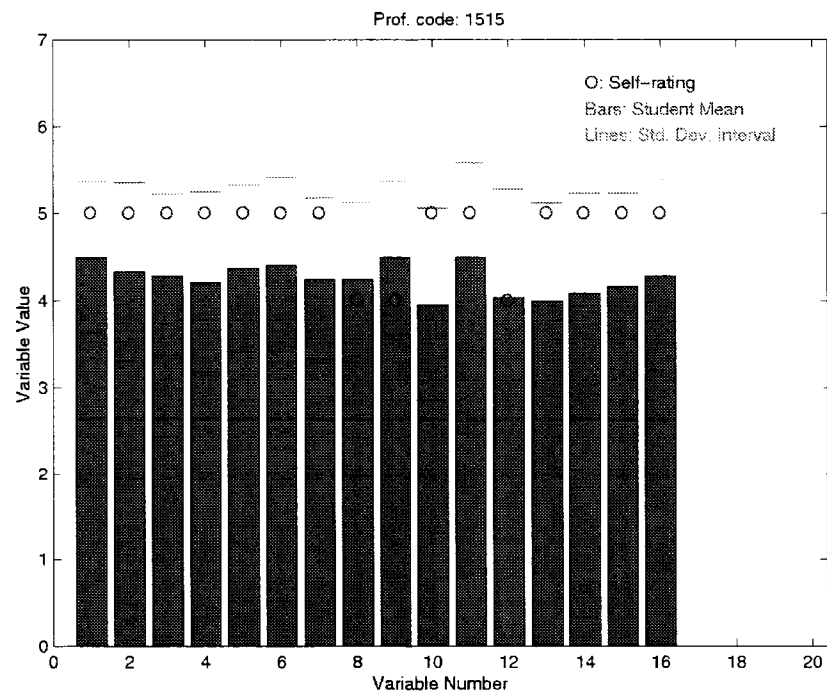
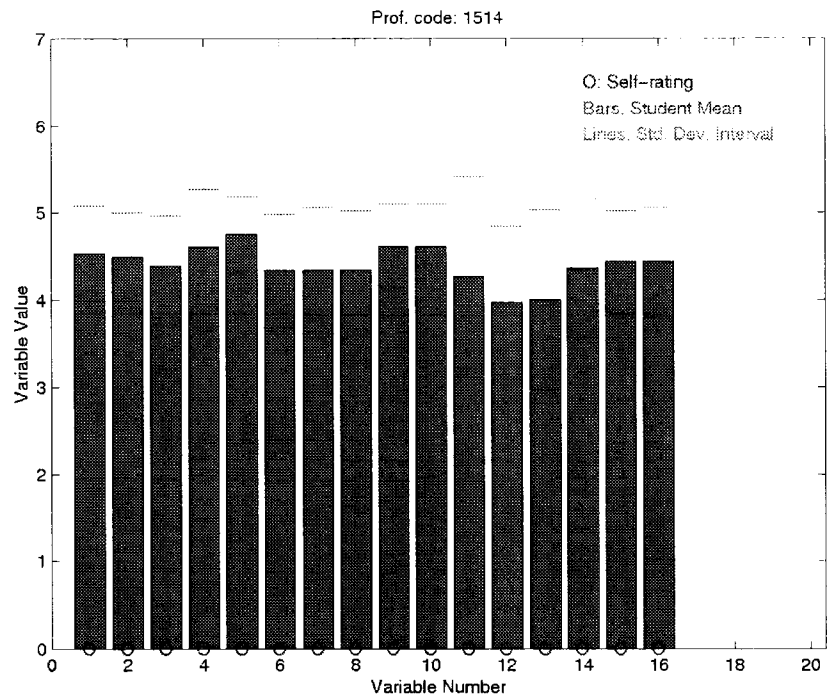


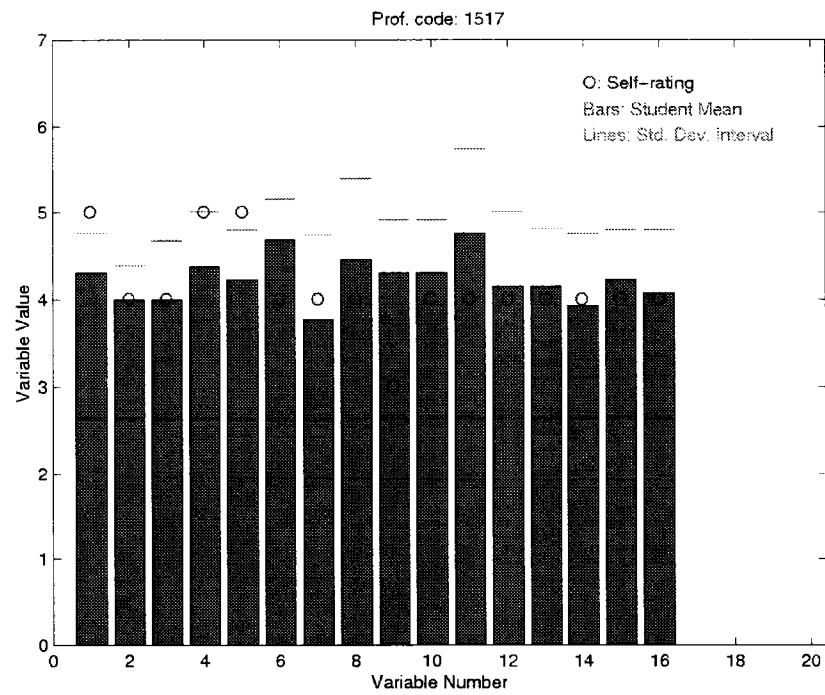
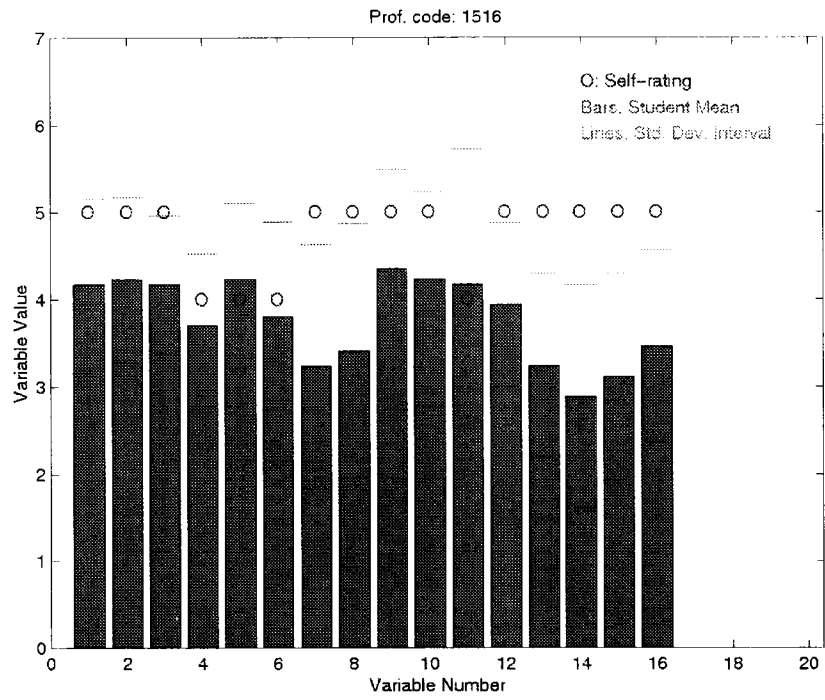
Prof. code: 1512

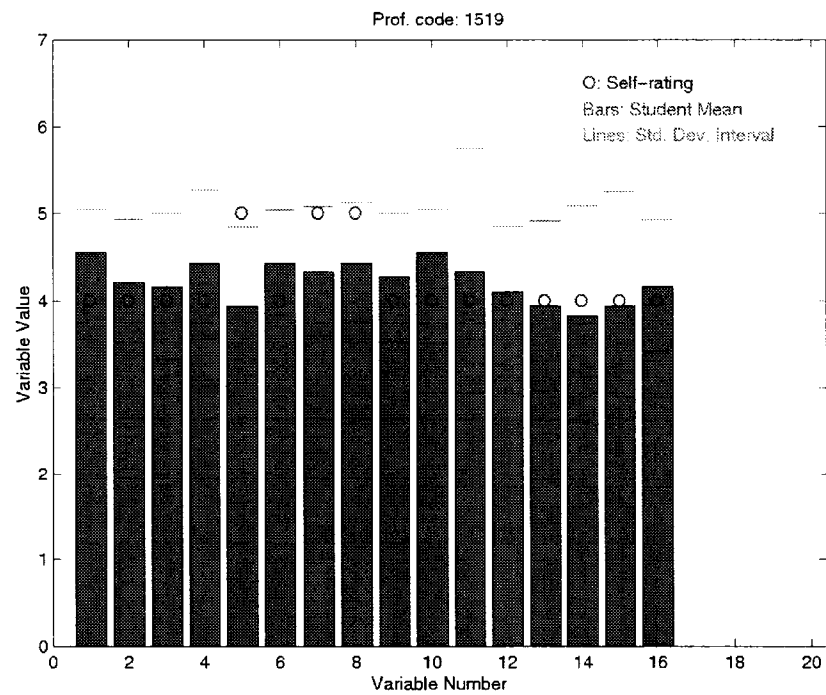
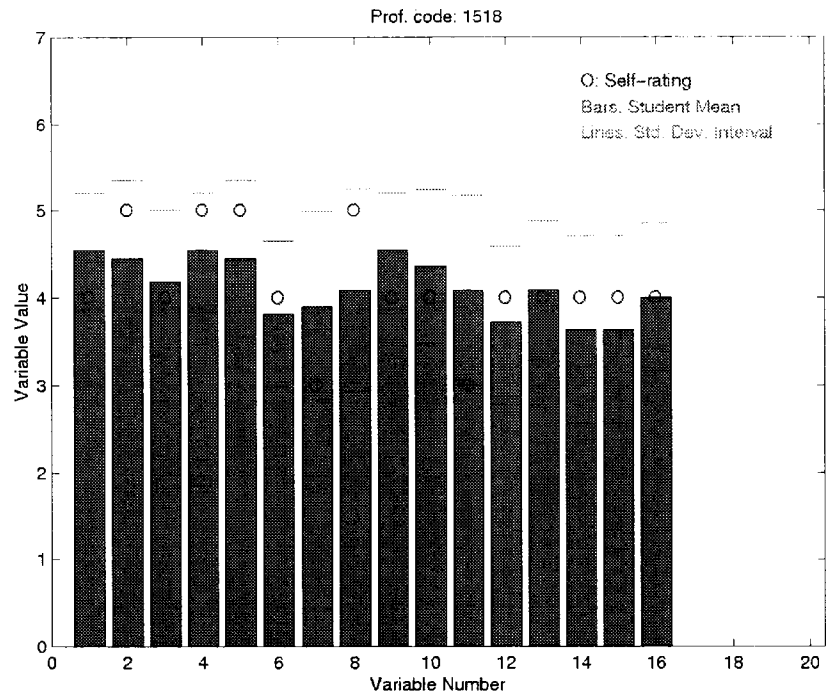


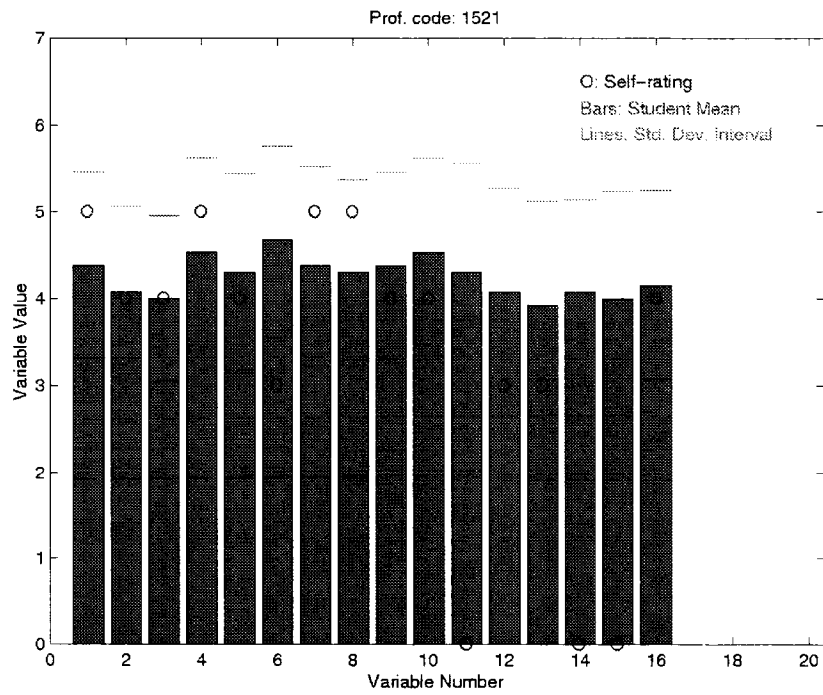
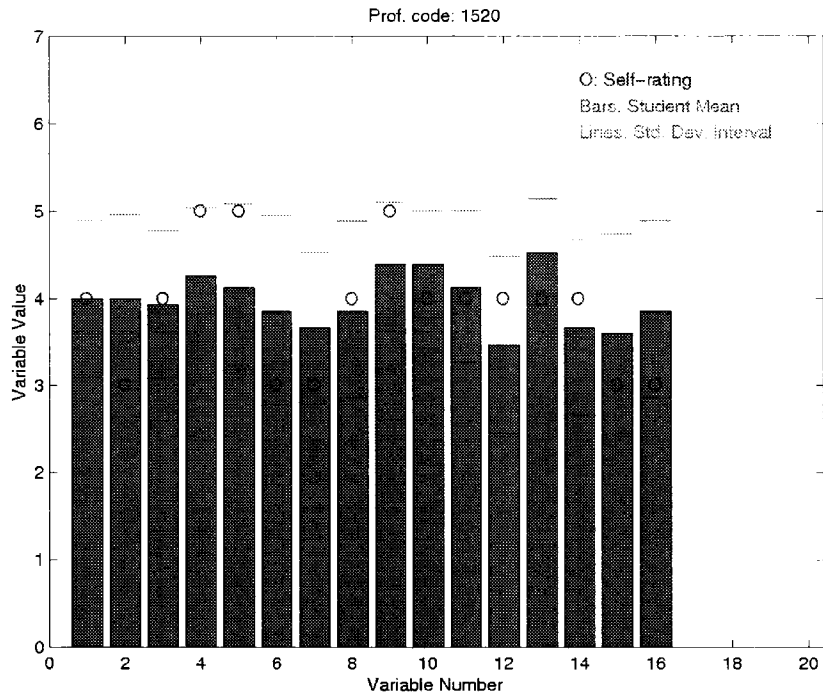
Prof. code: 1513



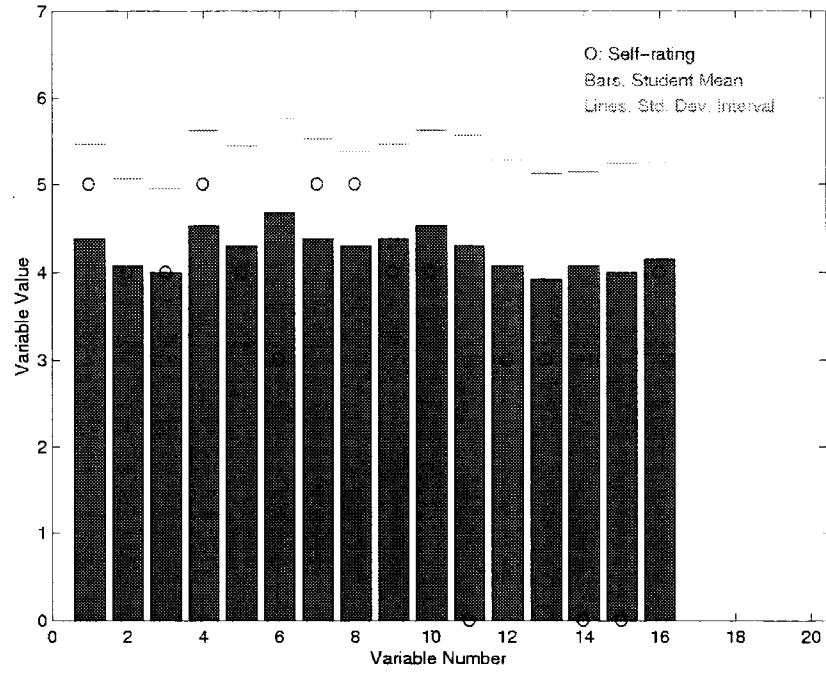


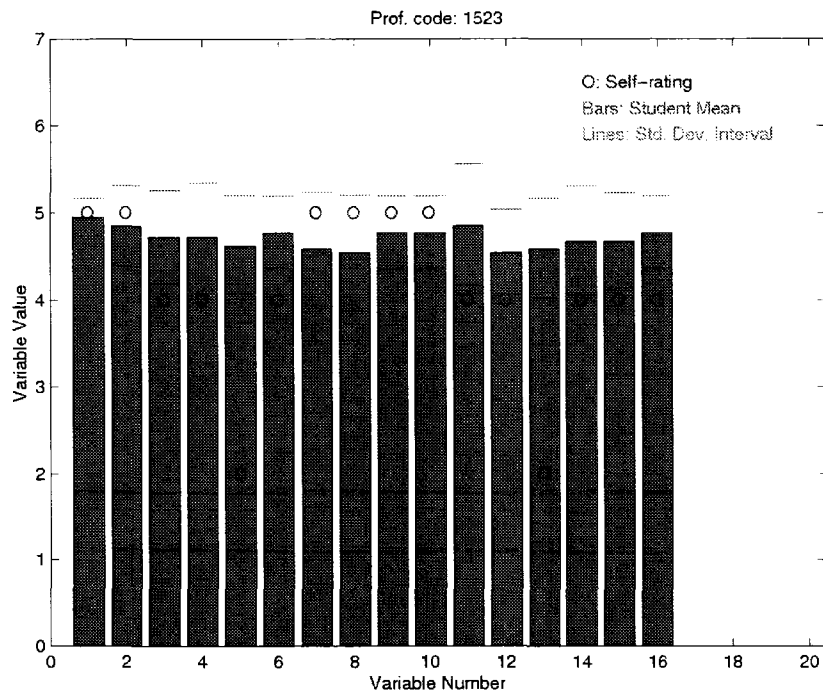
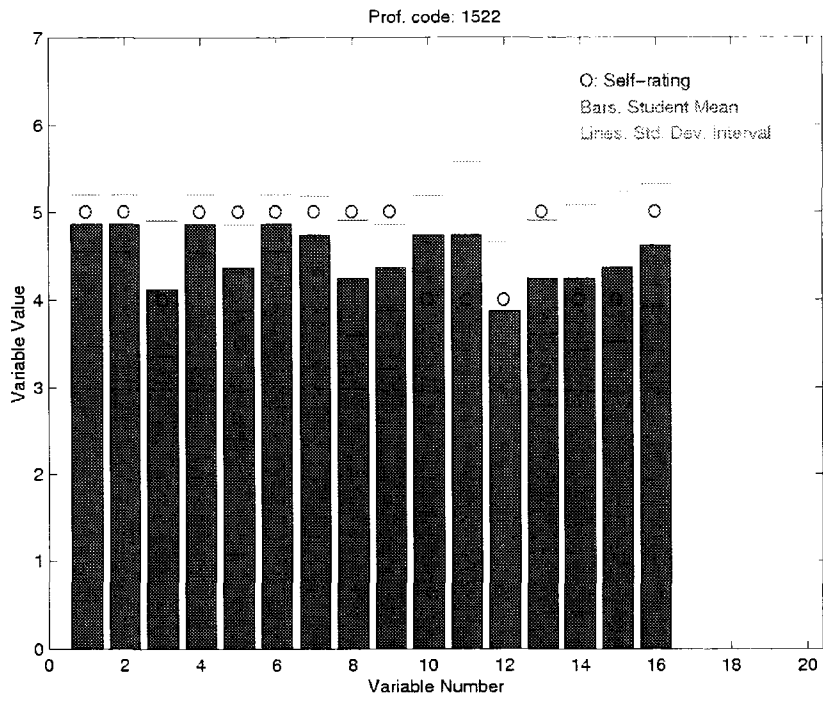


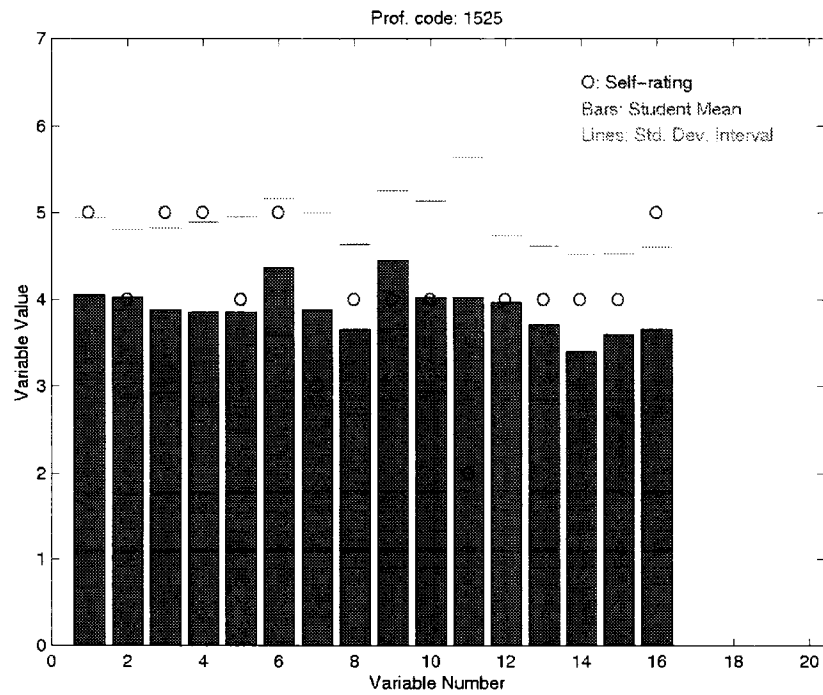
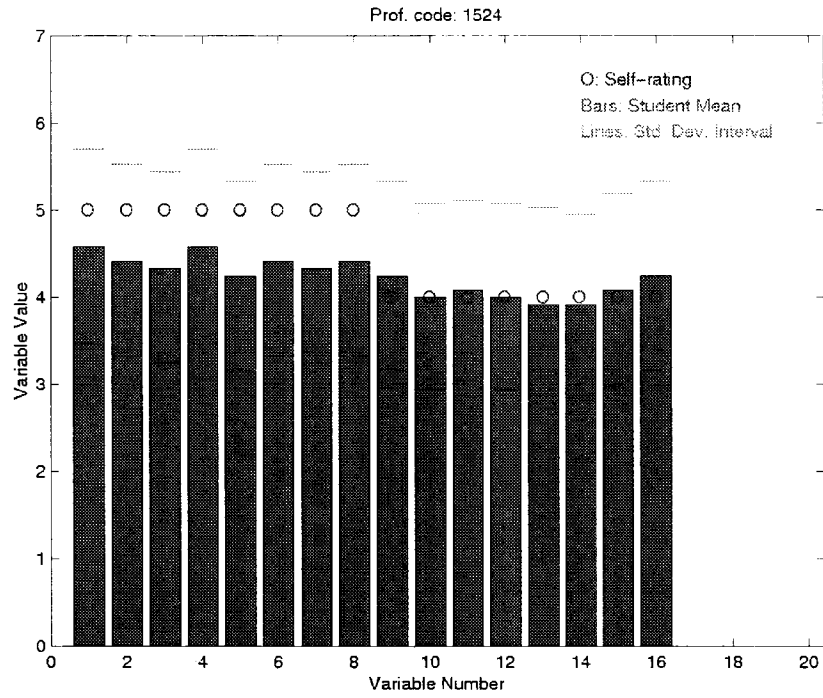


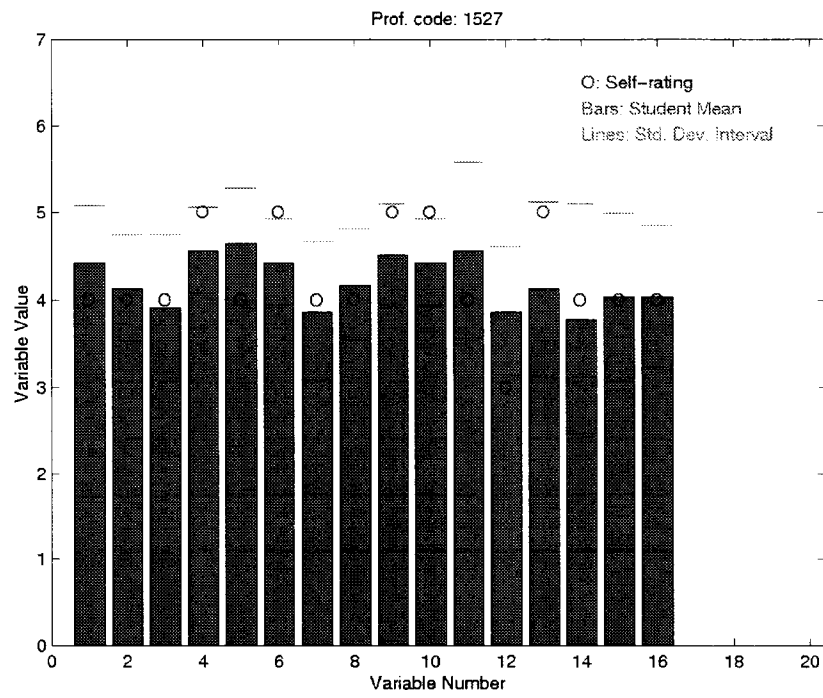
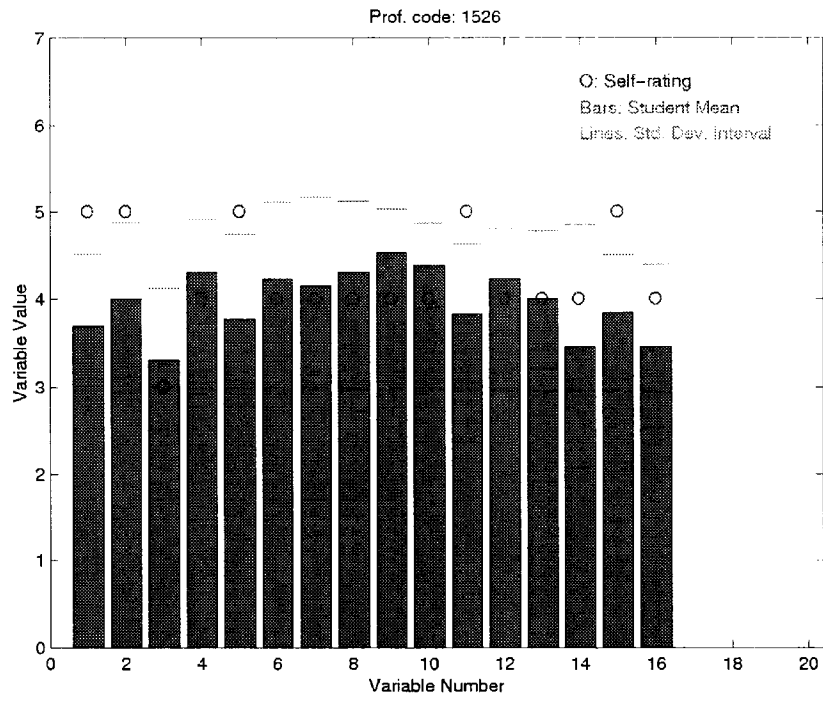


Prof. code: 1521

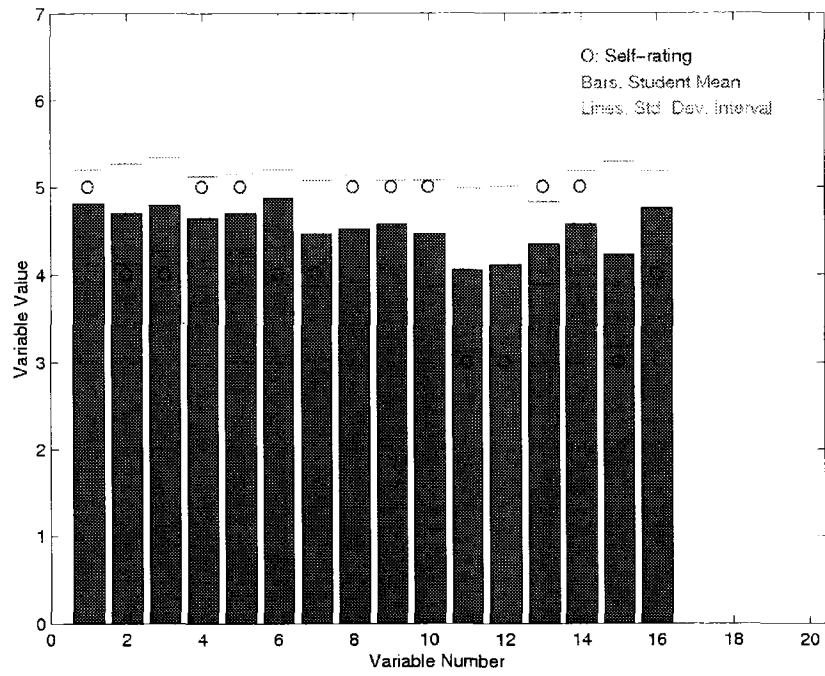




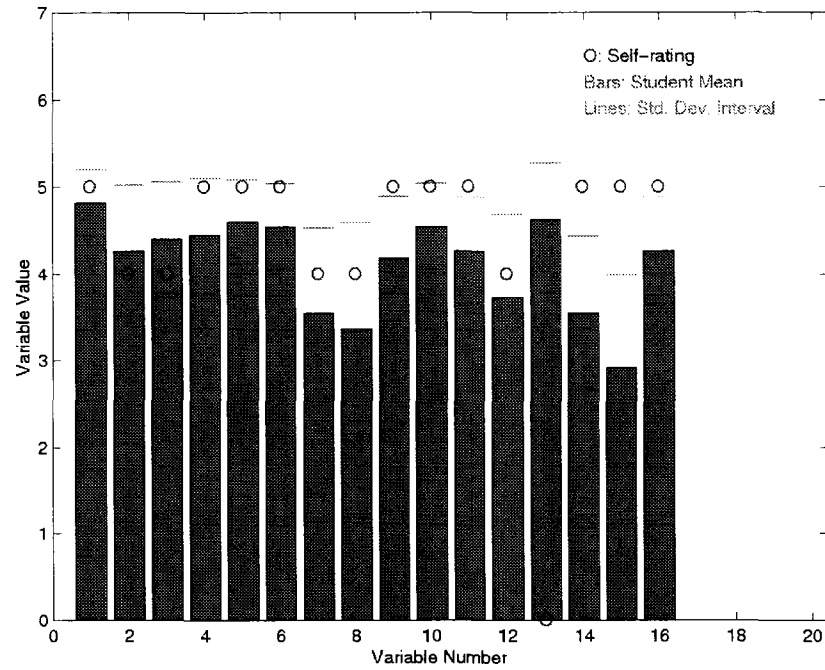


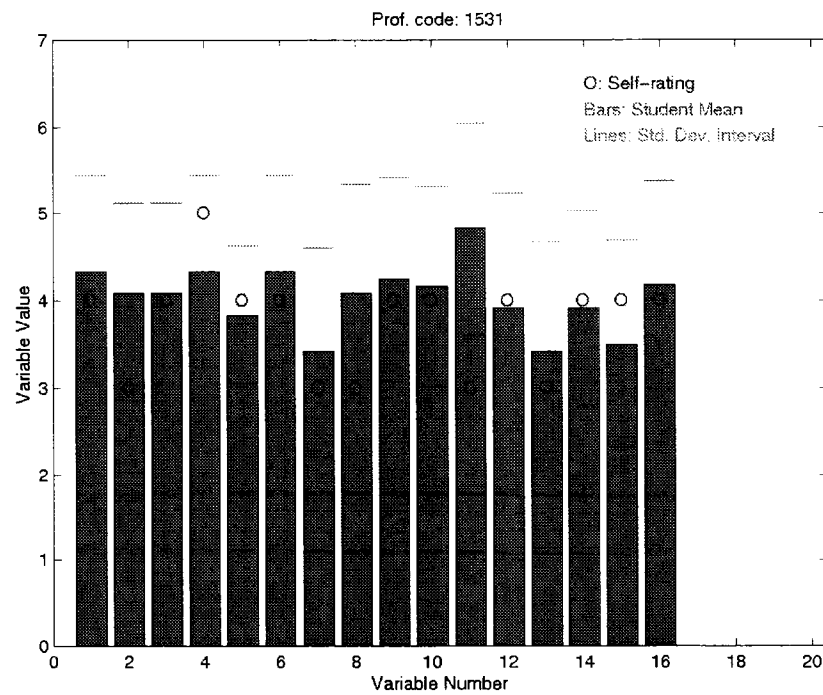
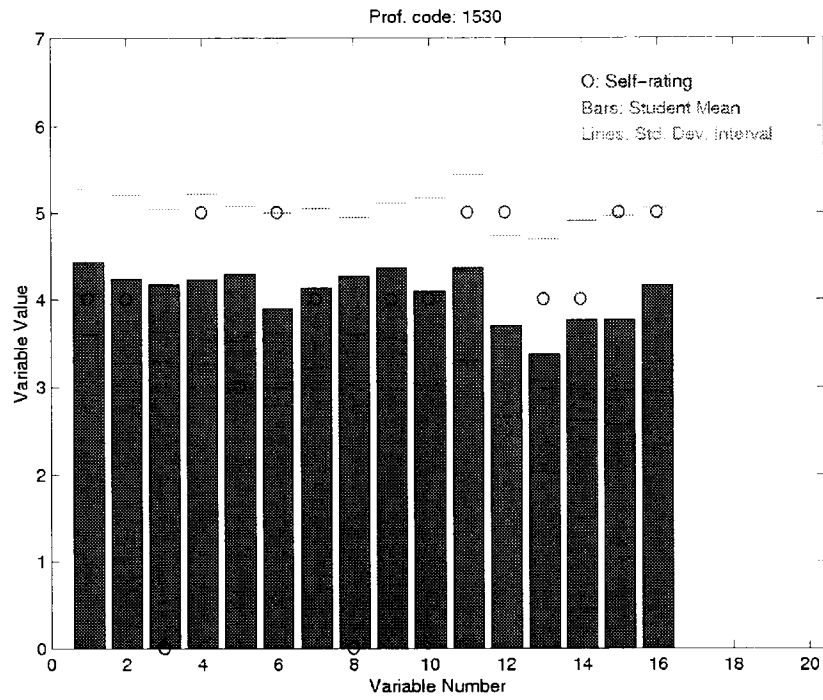


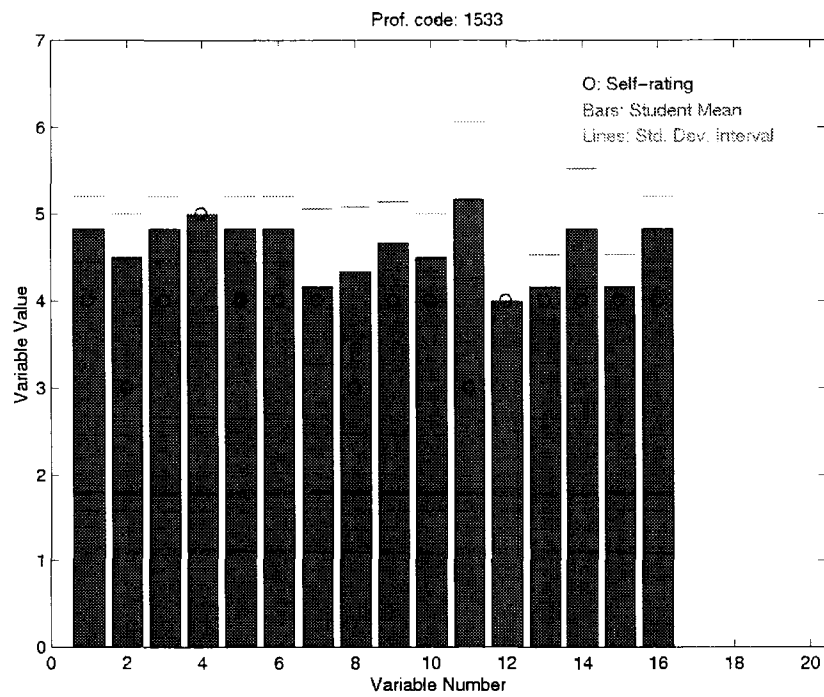
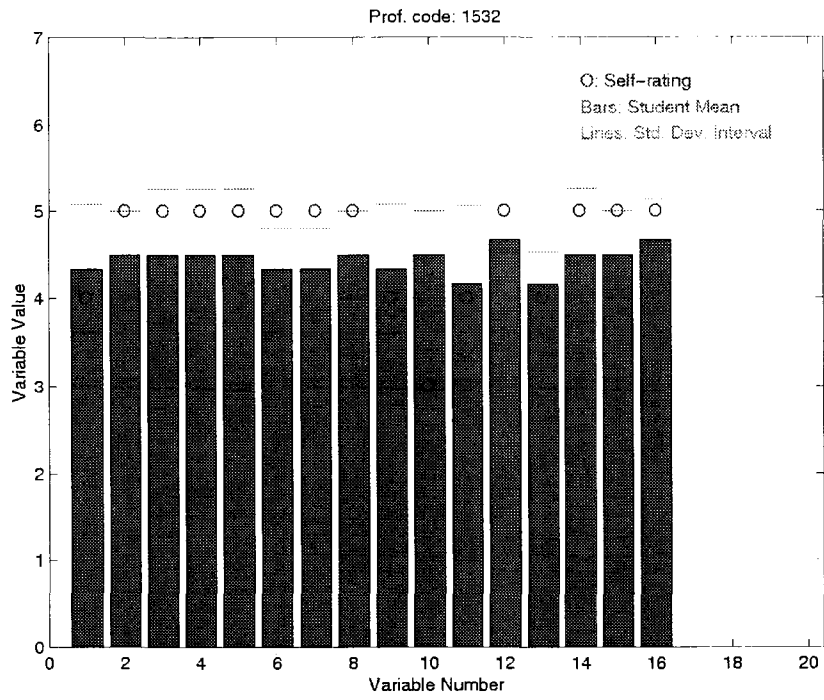
Prof. code: 1528

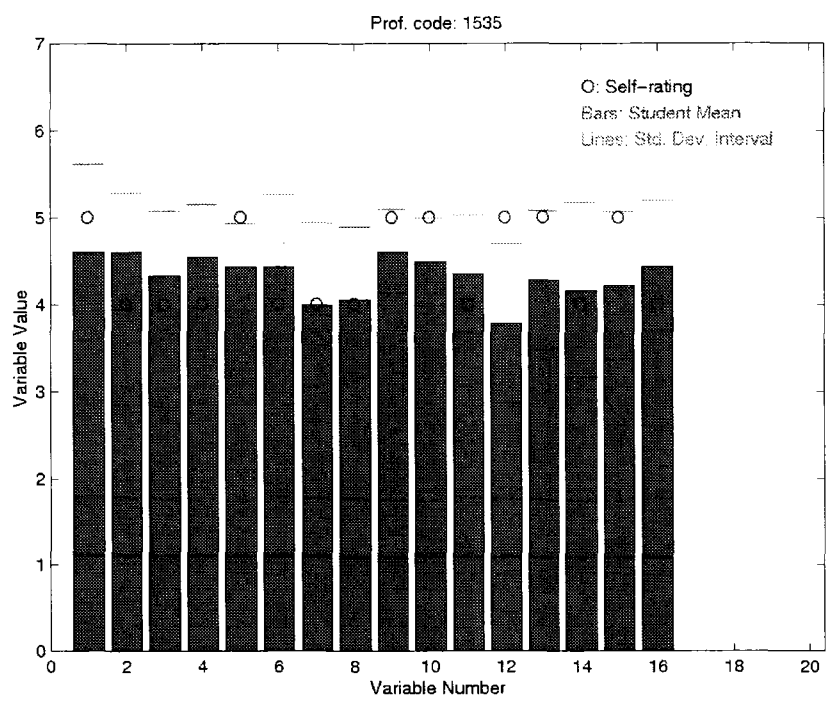
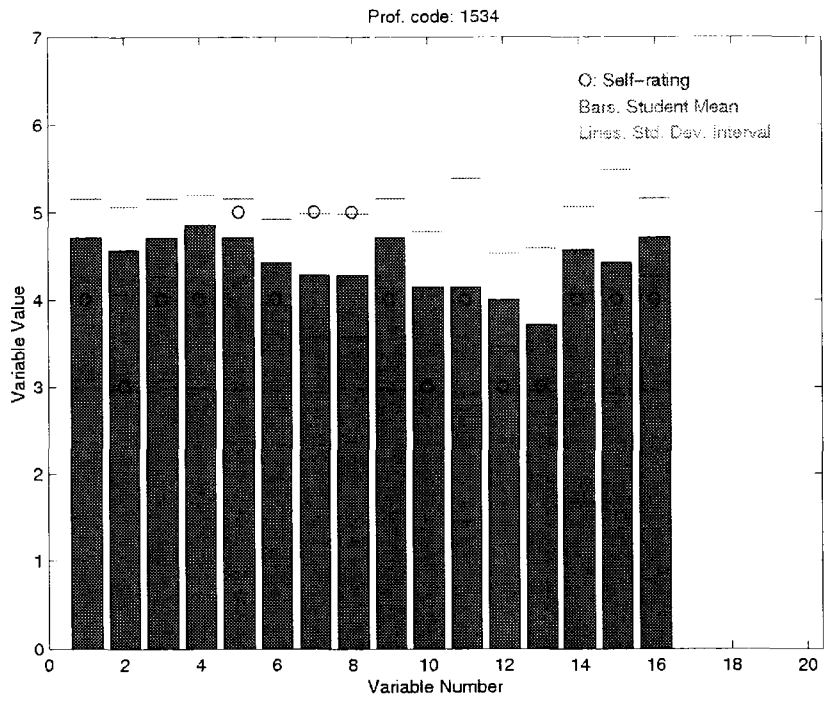


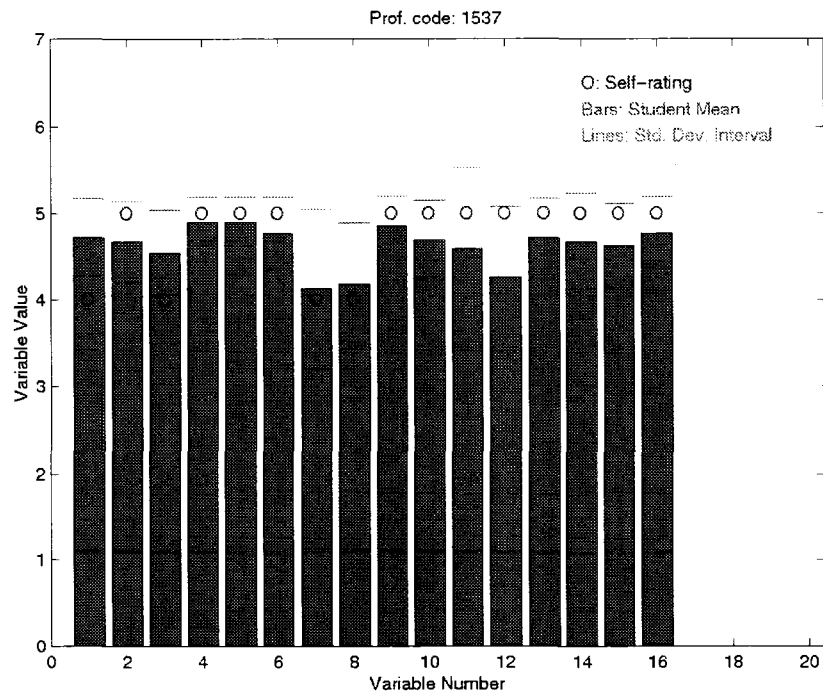
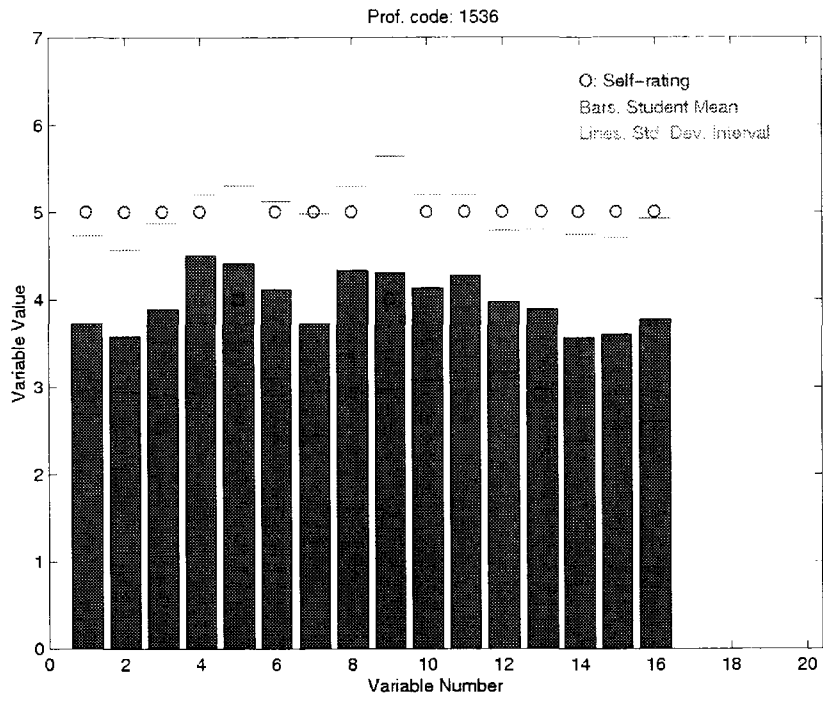
Prof. code: 1529

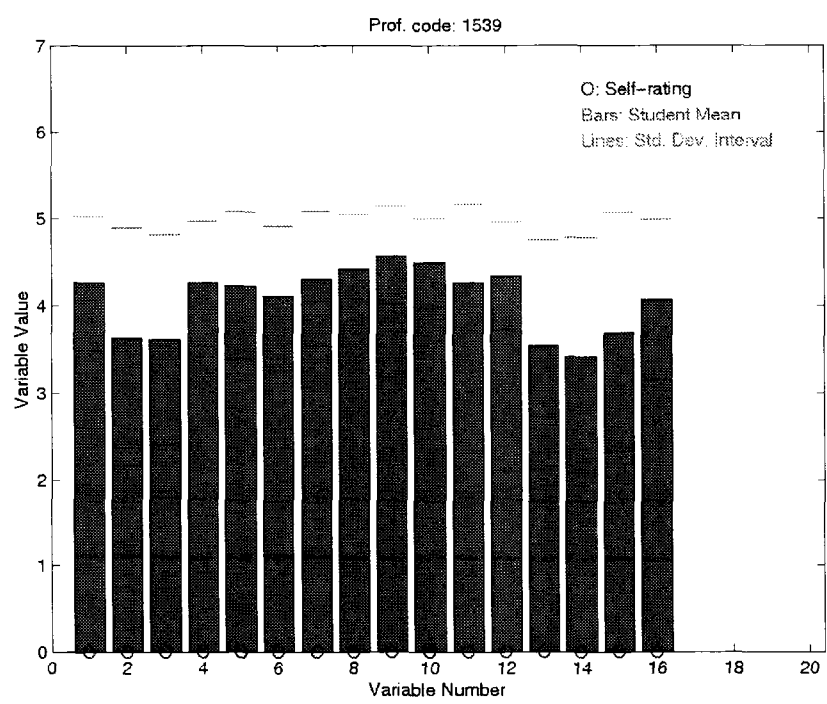
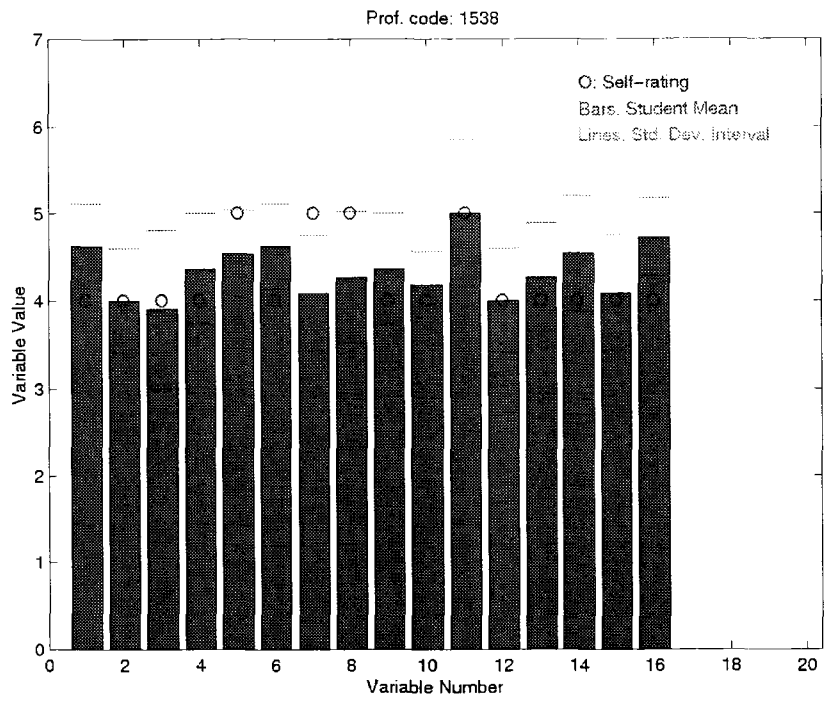


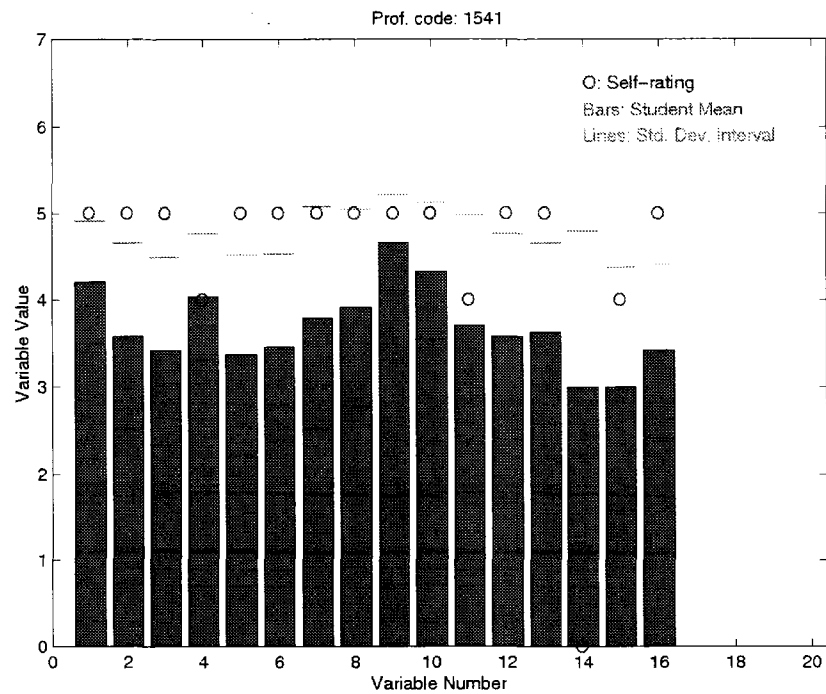
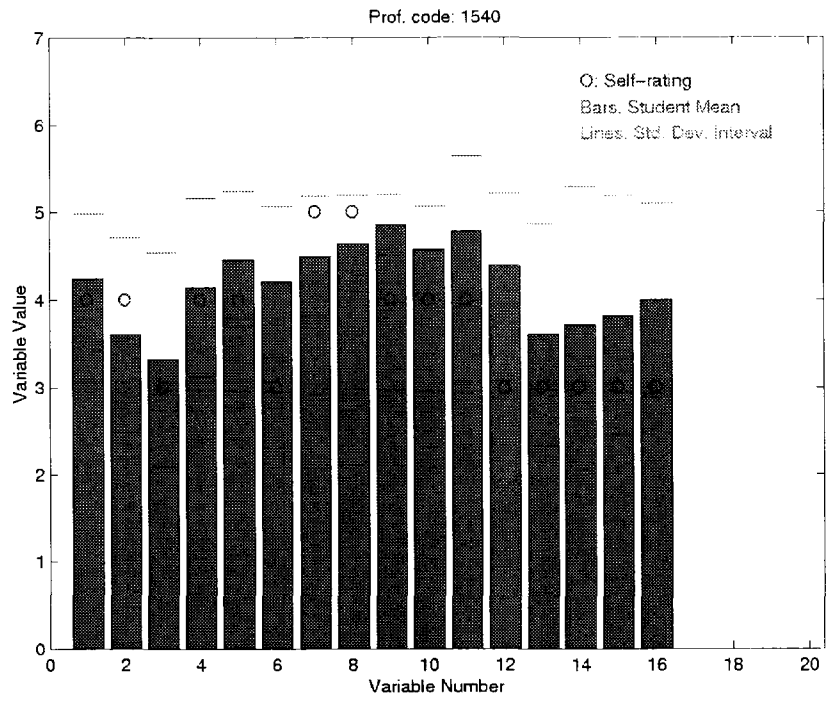


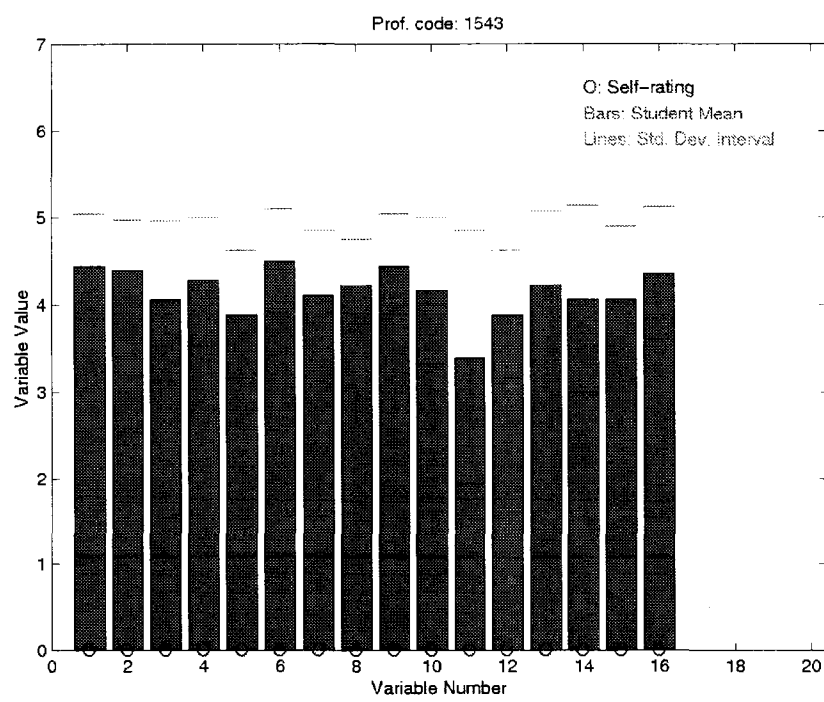
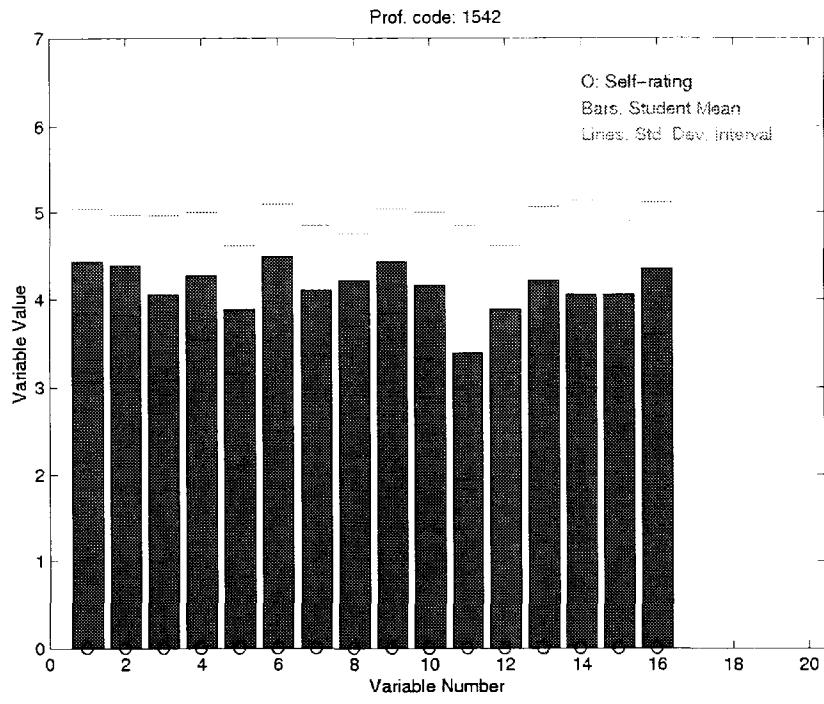


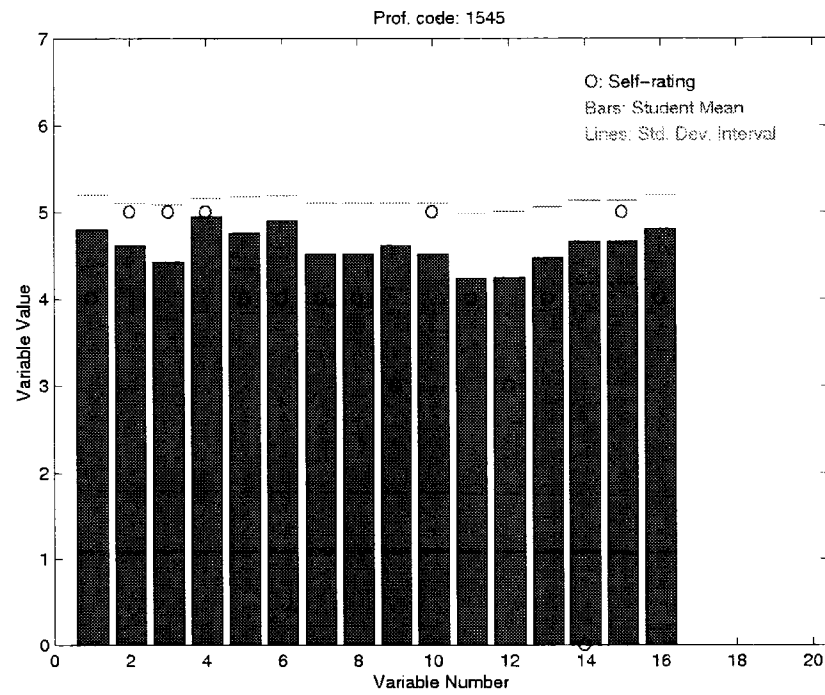
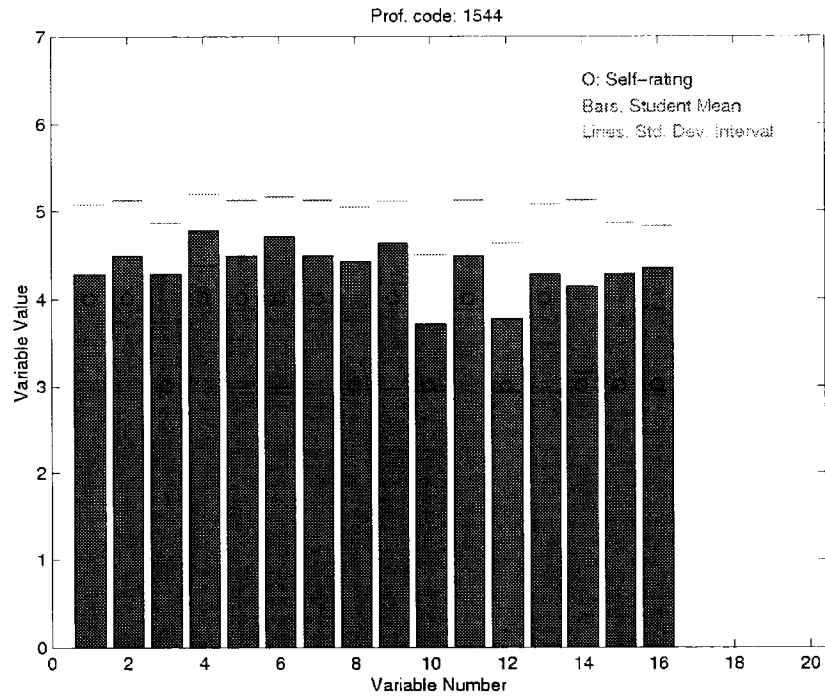


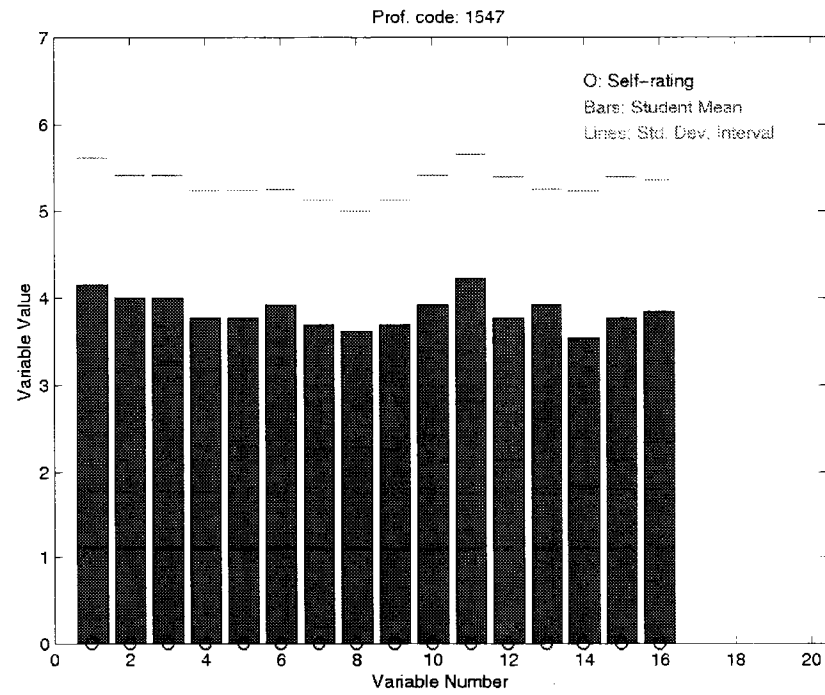
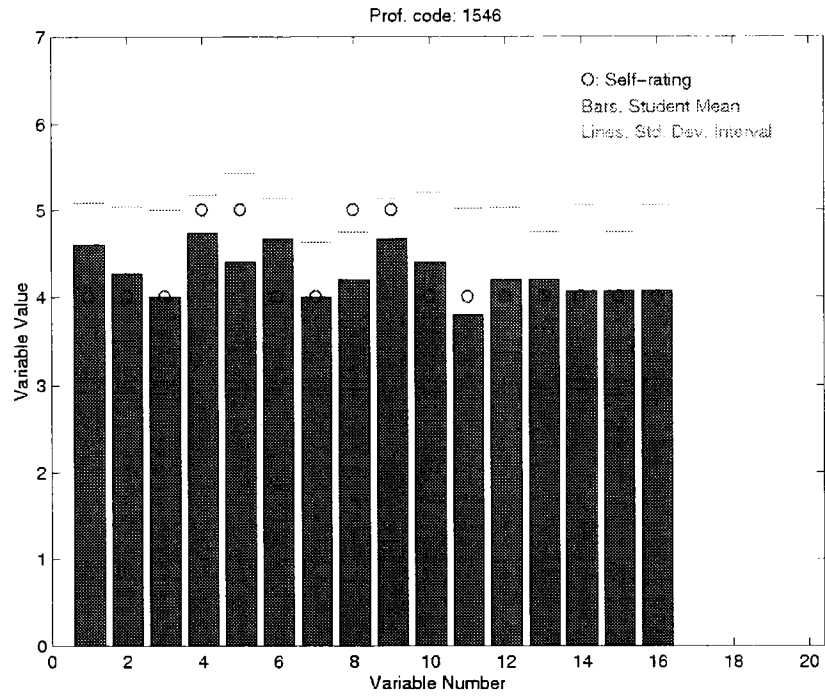


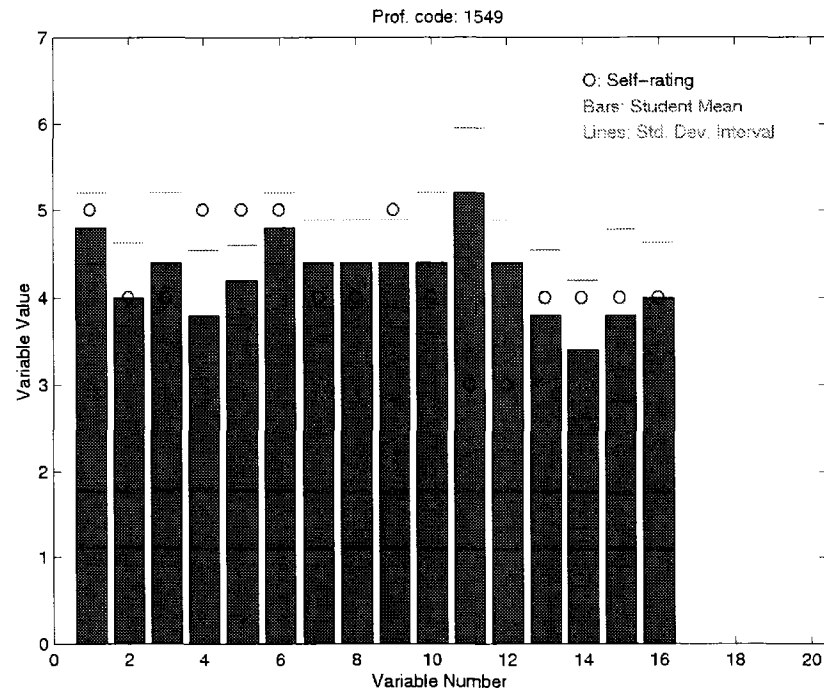
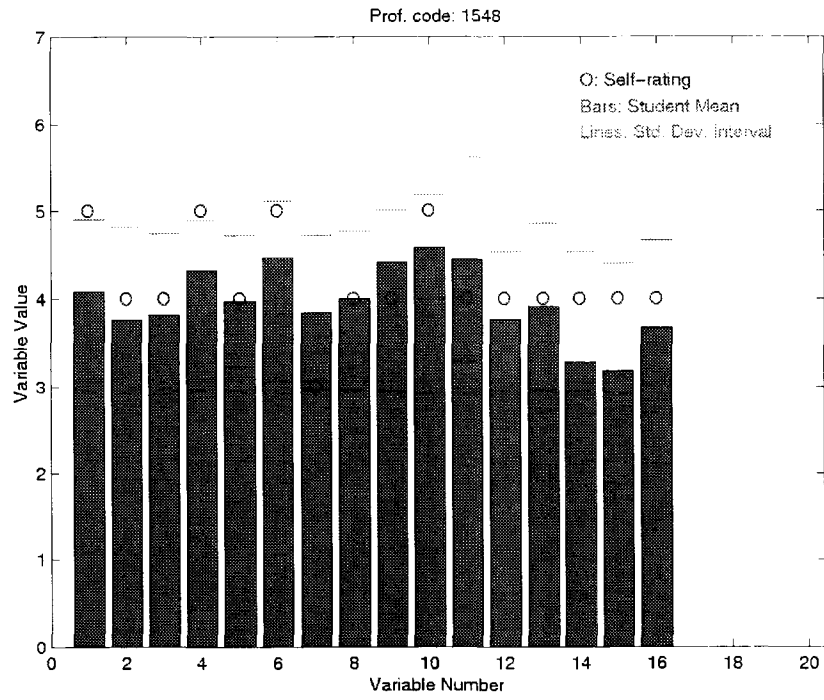


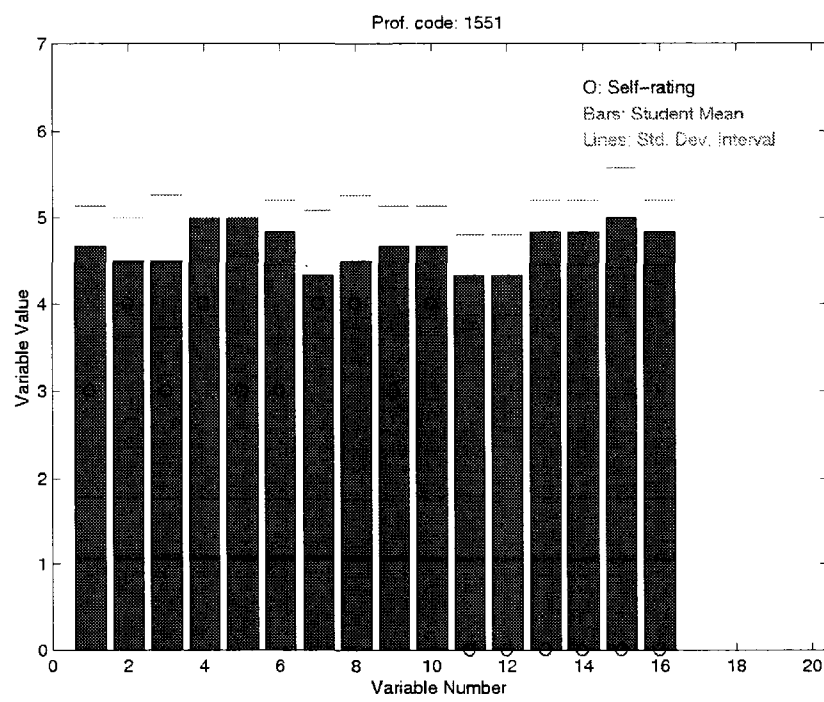
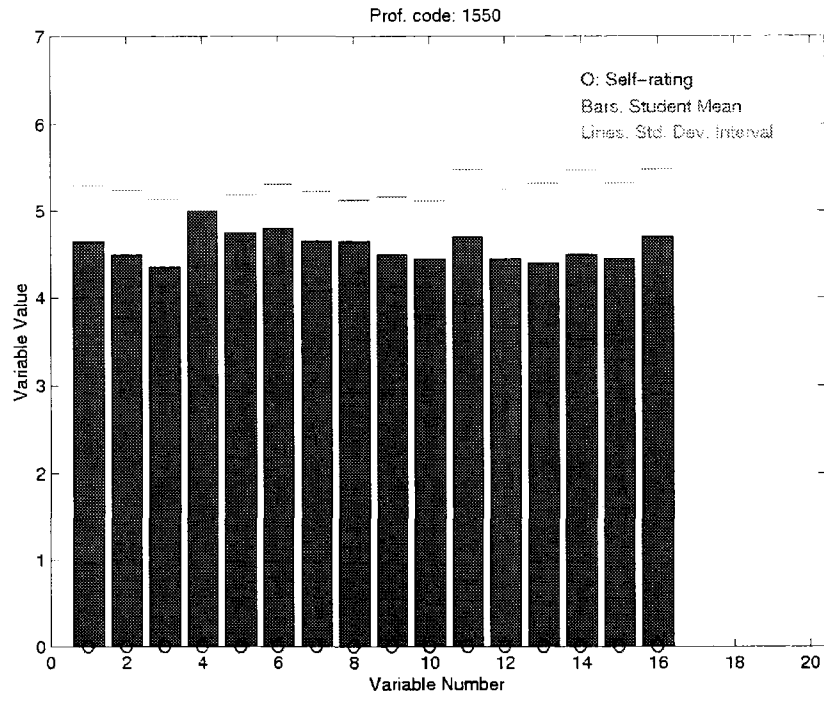


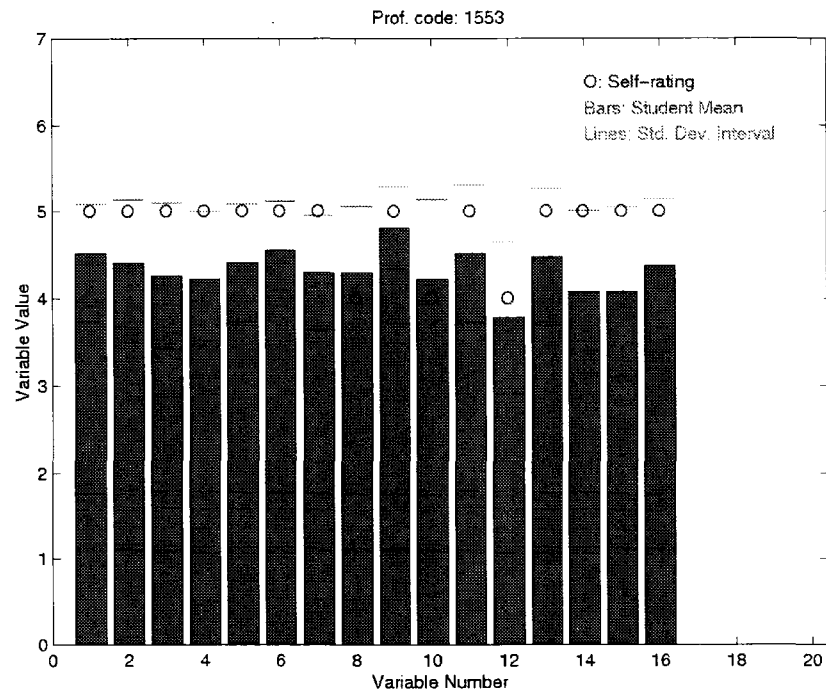
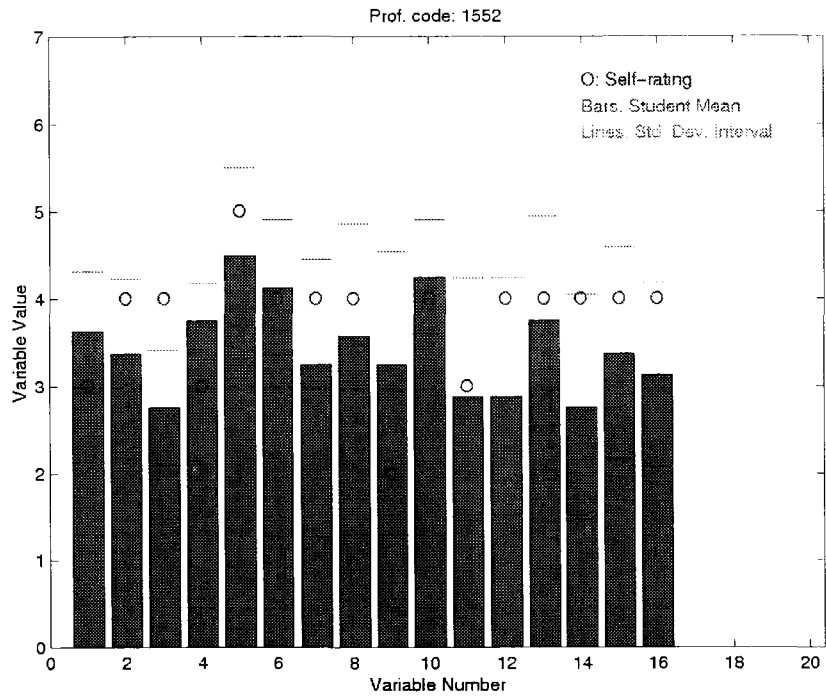


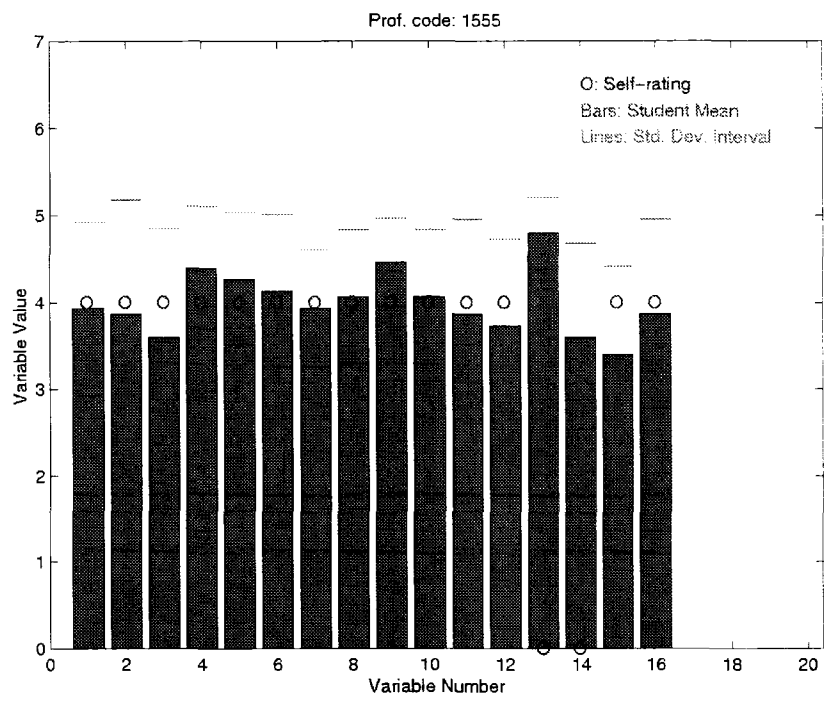
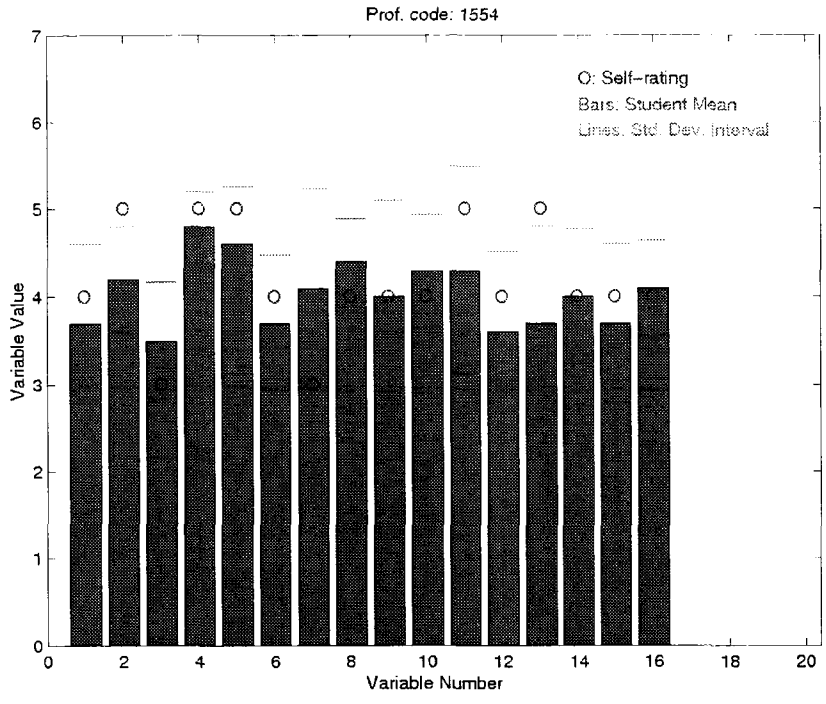


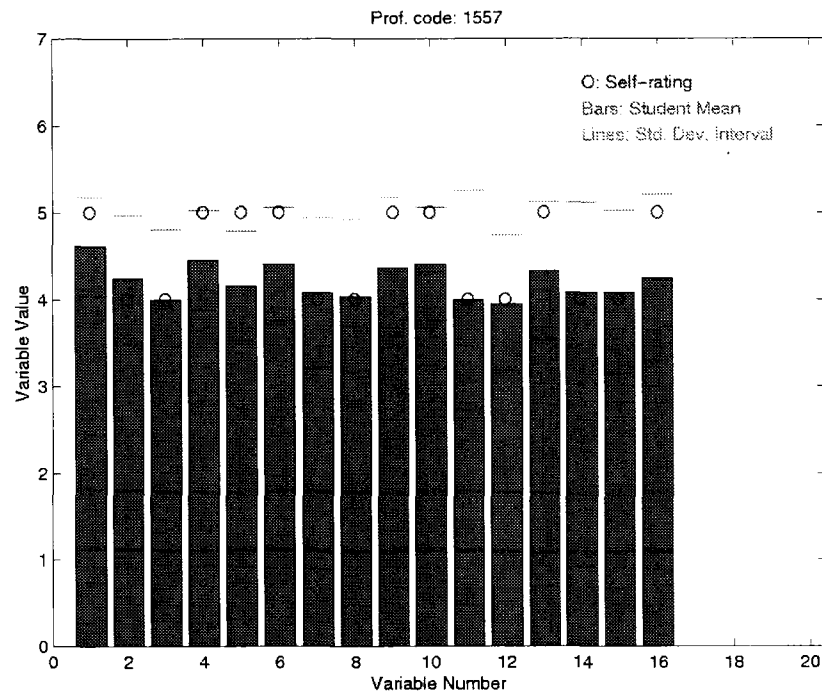
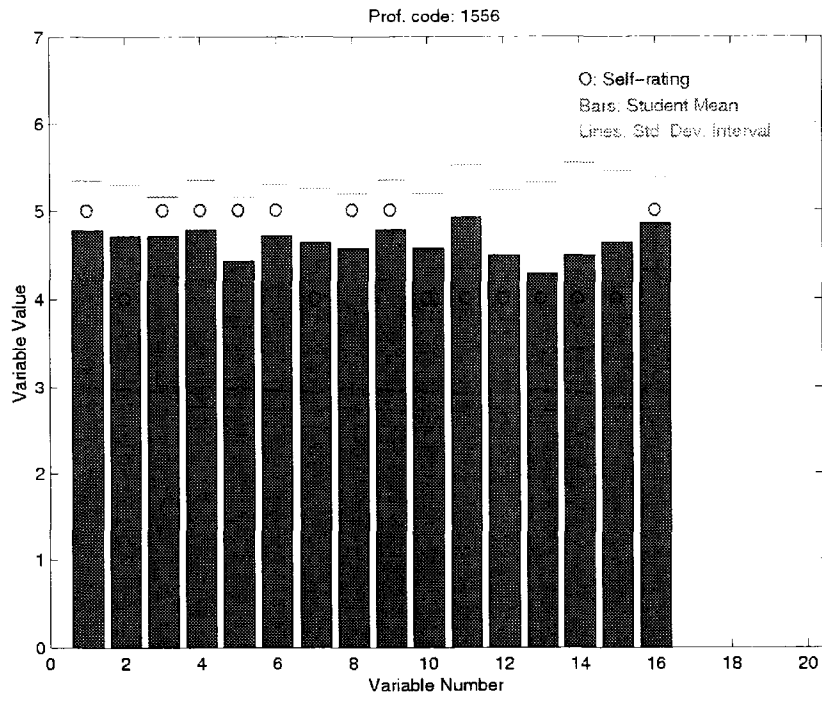


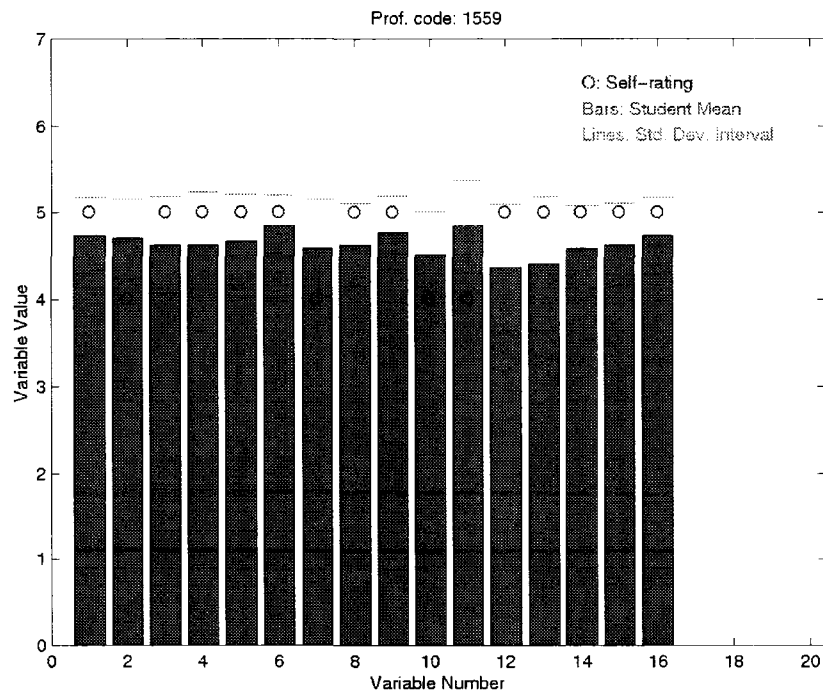
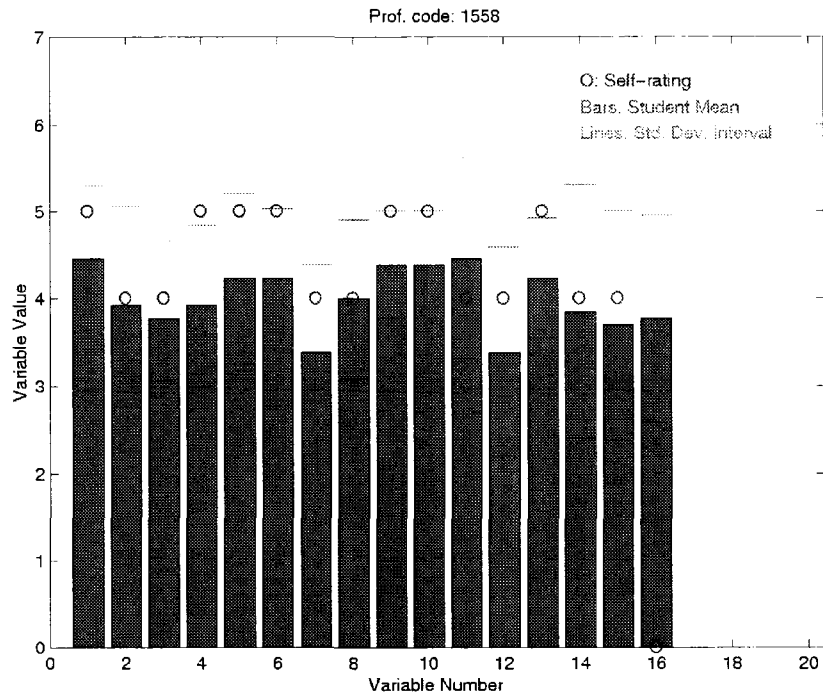


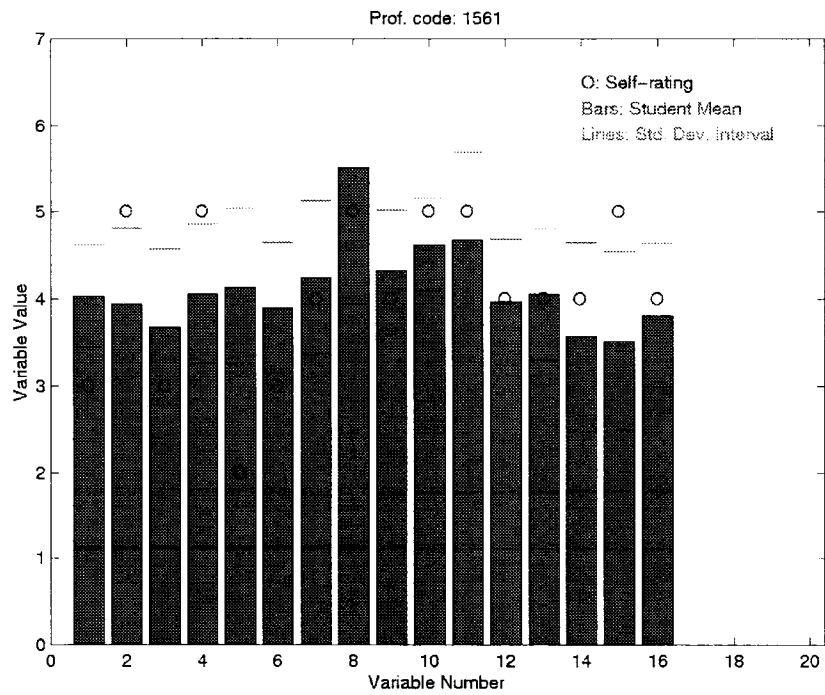
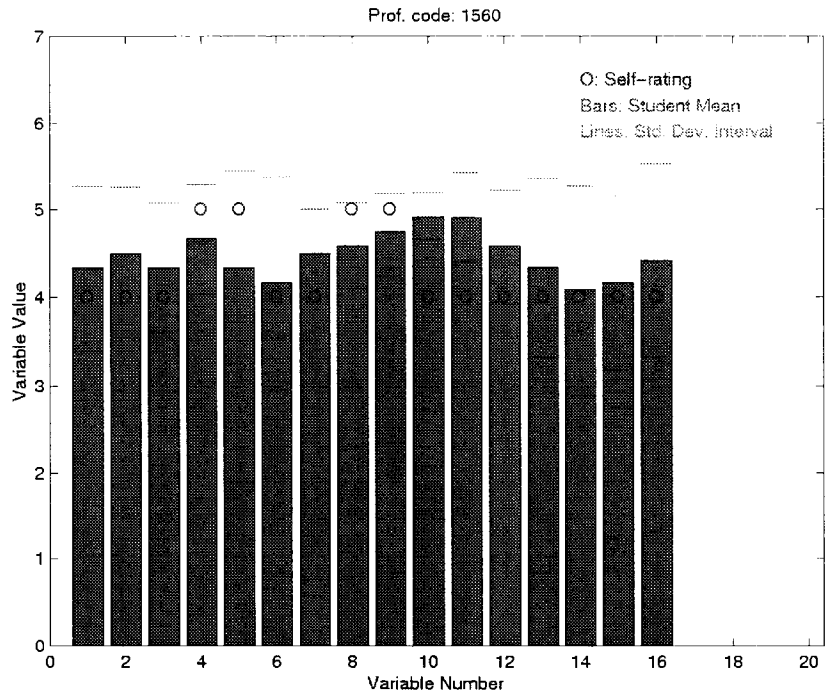


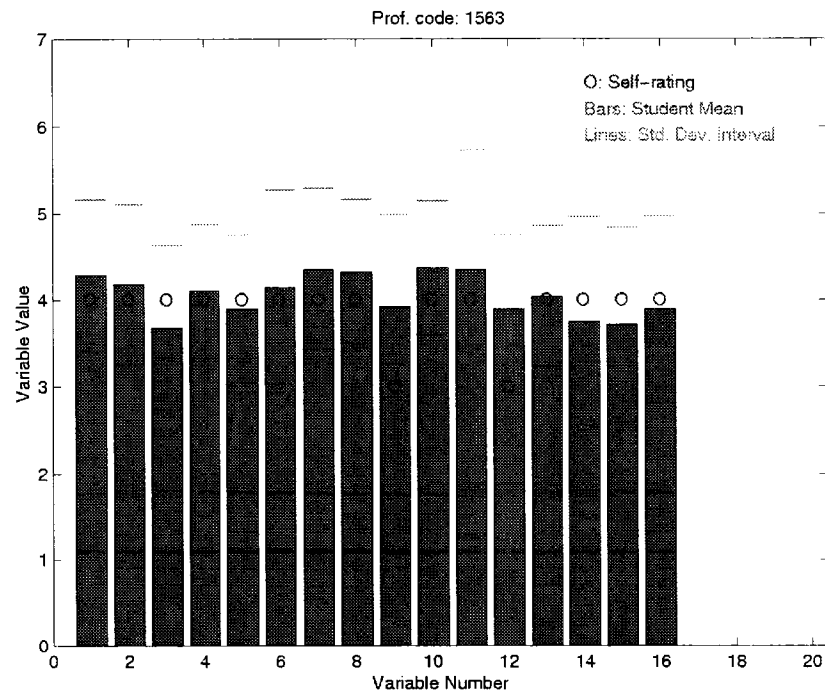
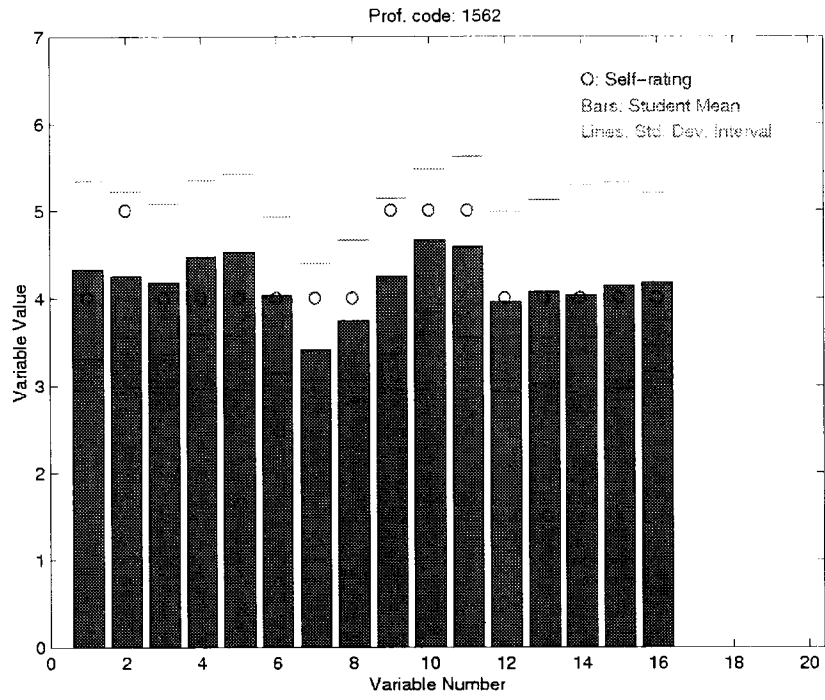




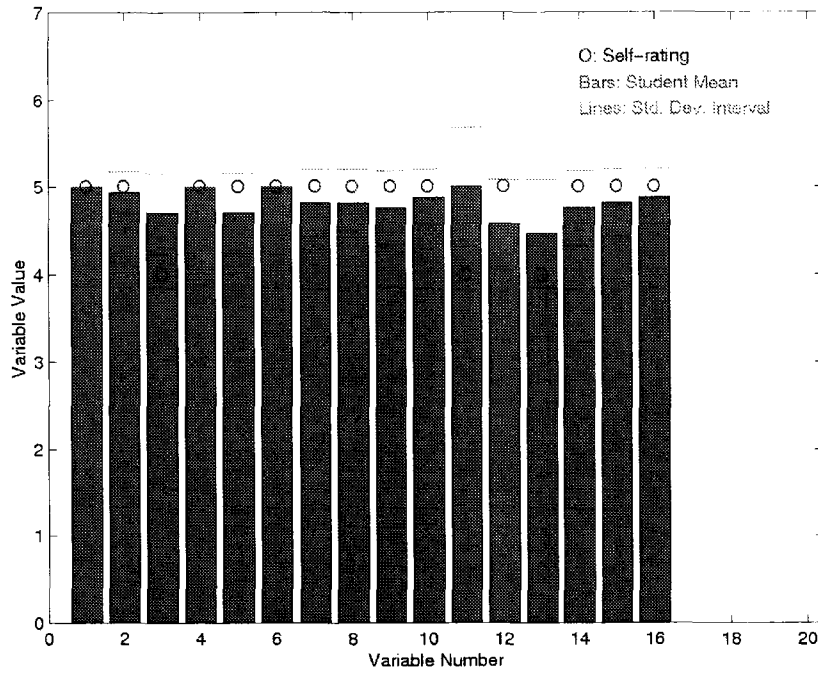




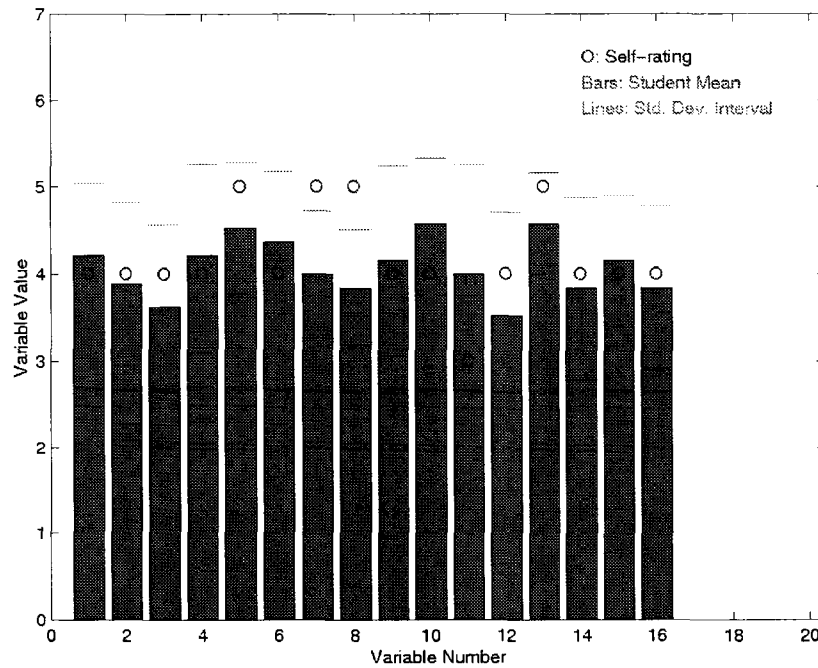


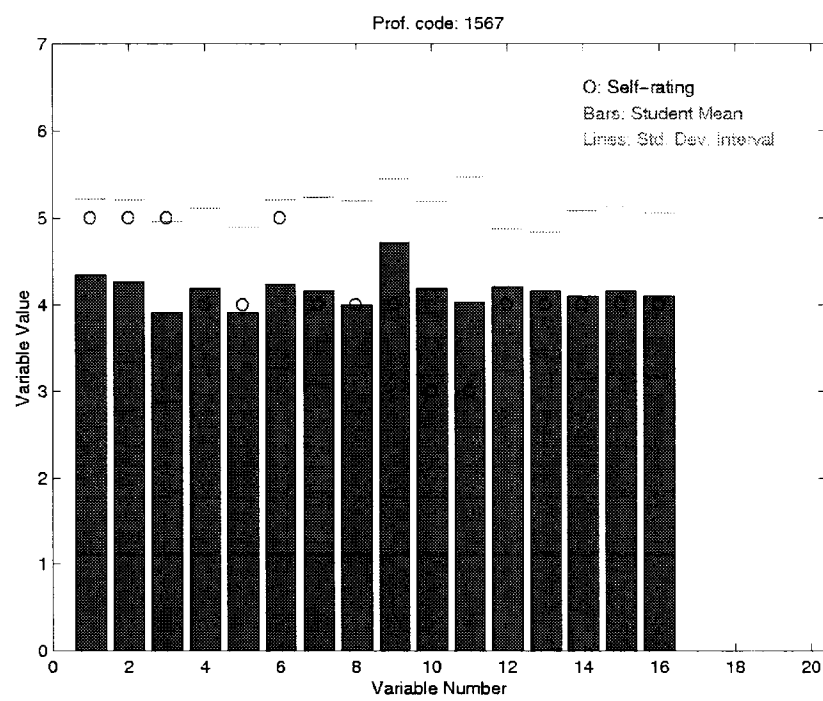
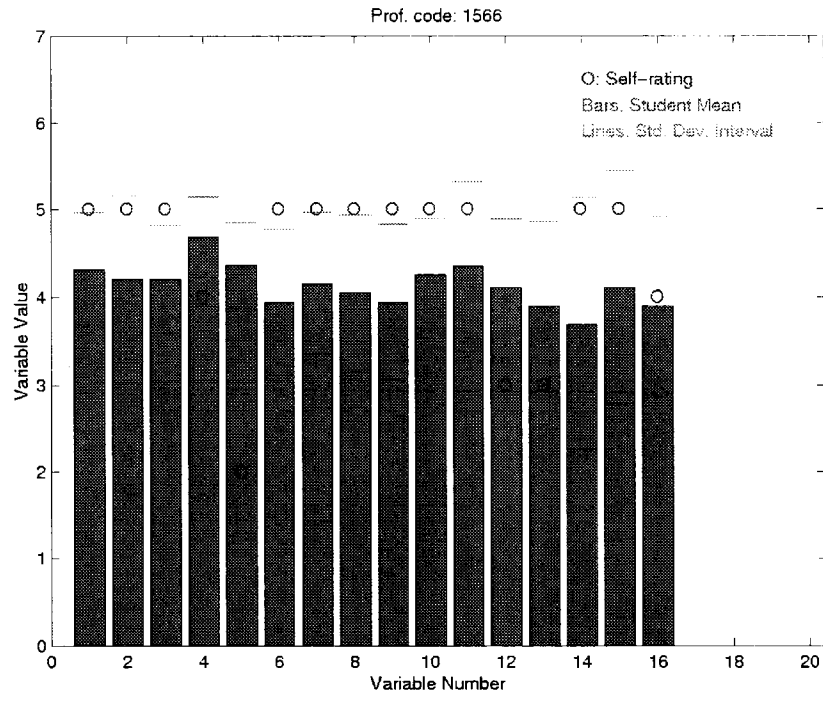


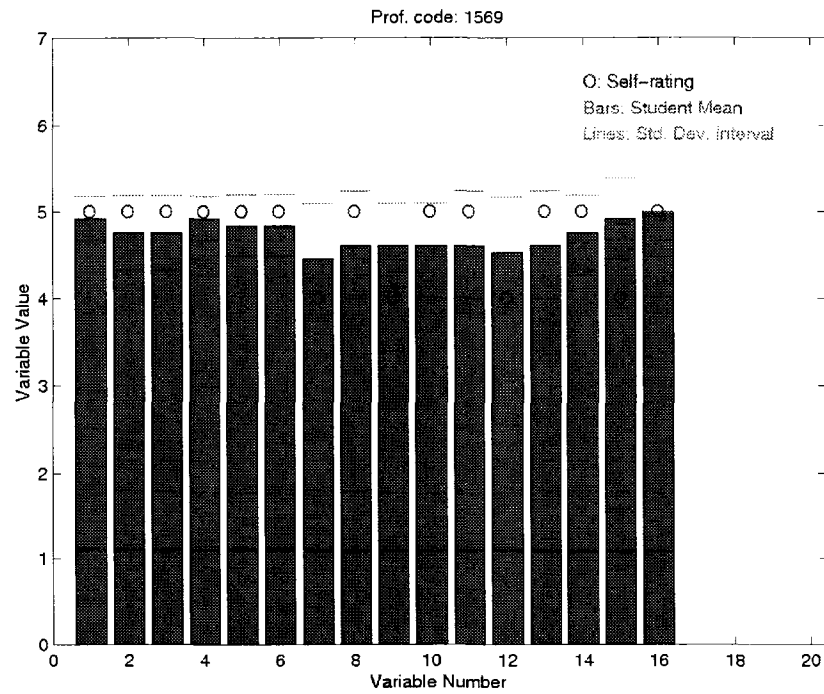
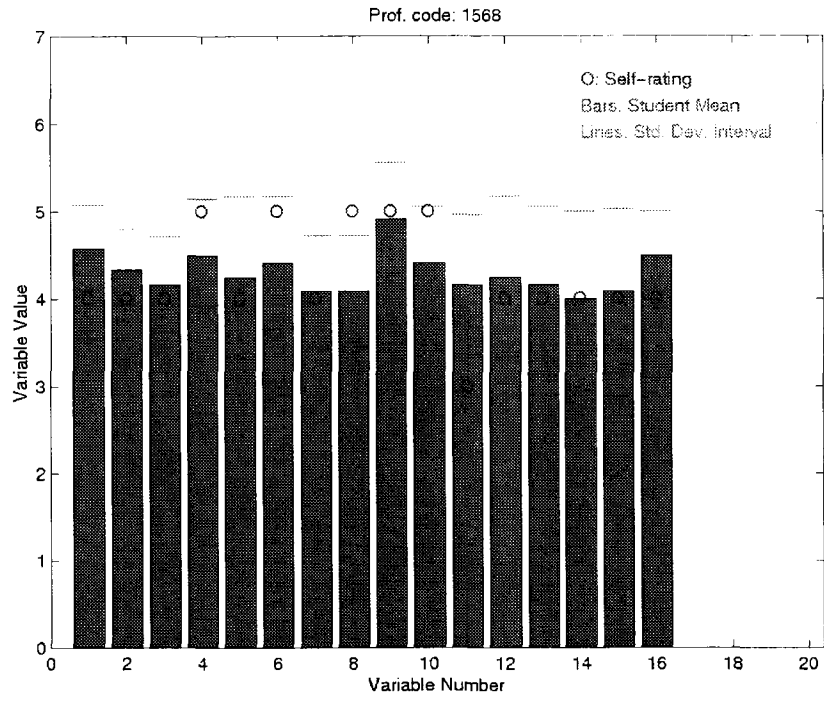
Prof. code: 1564

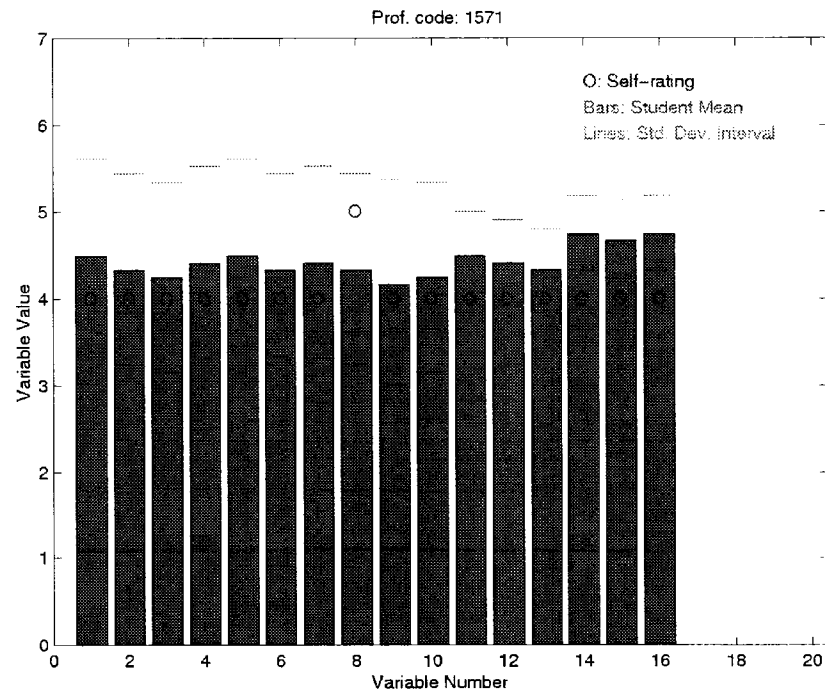
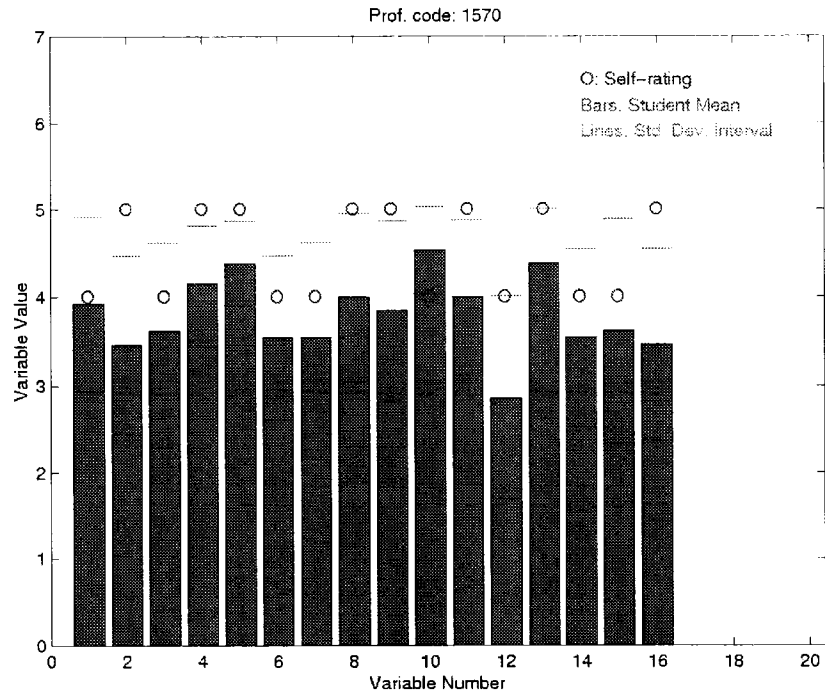


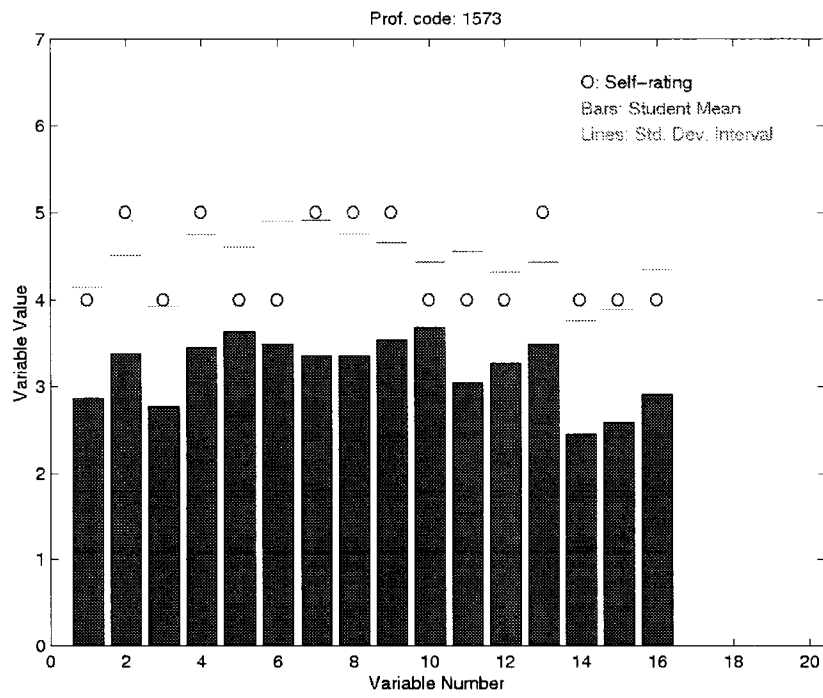
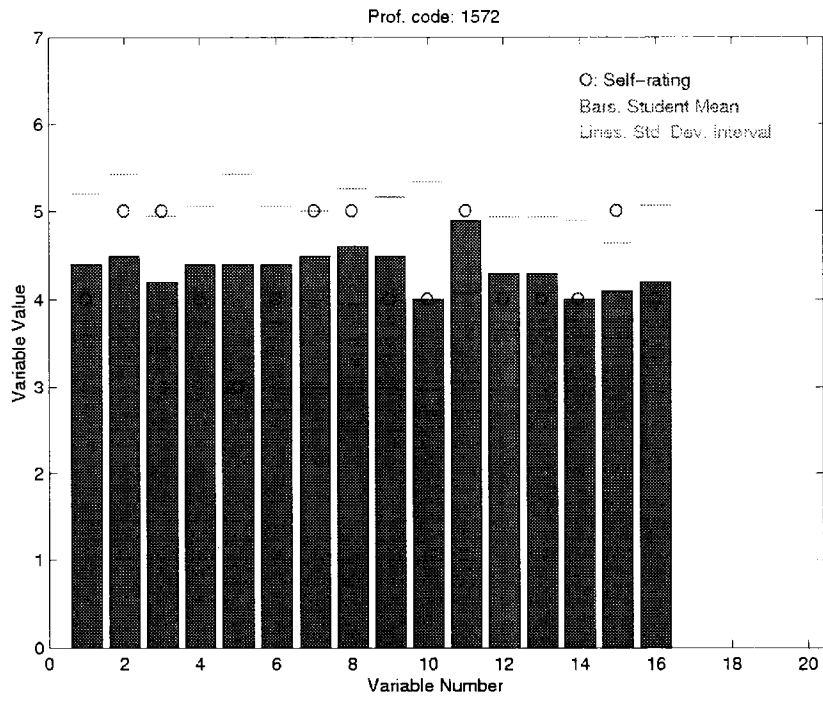
Prof. code: 1565

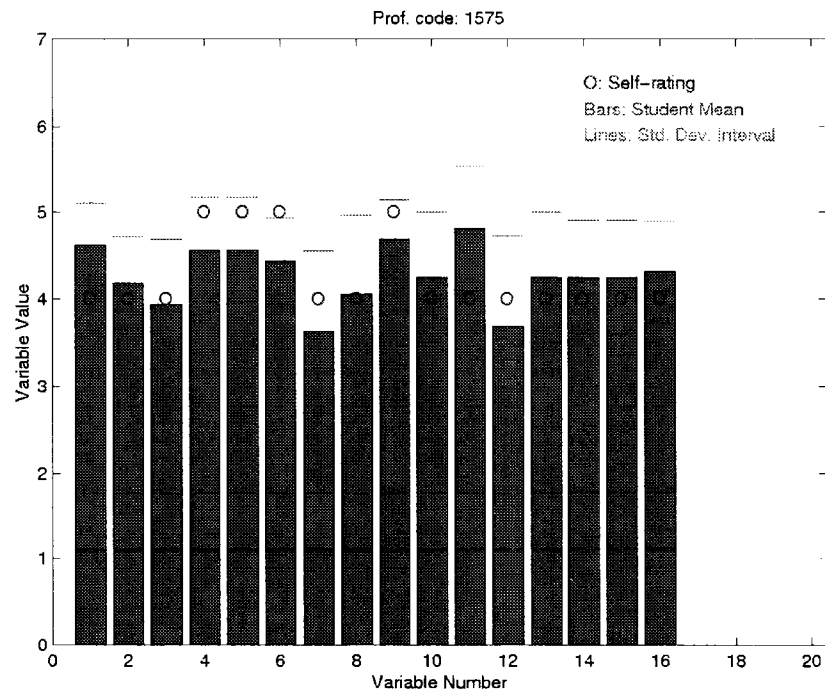
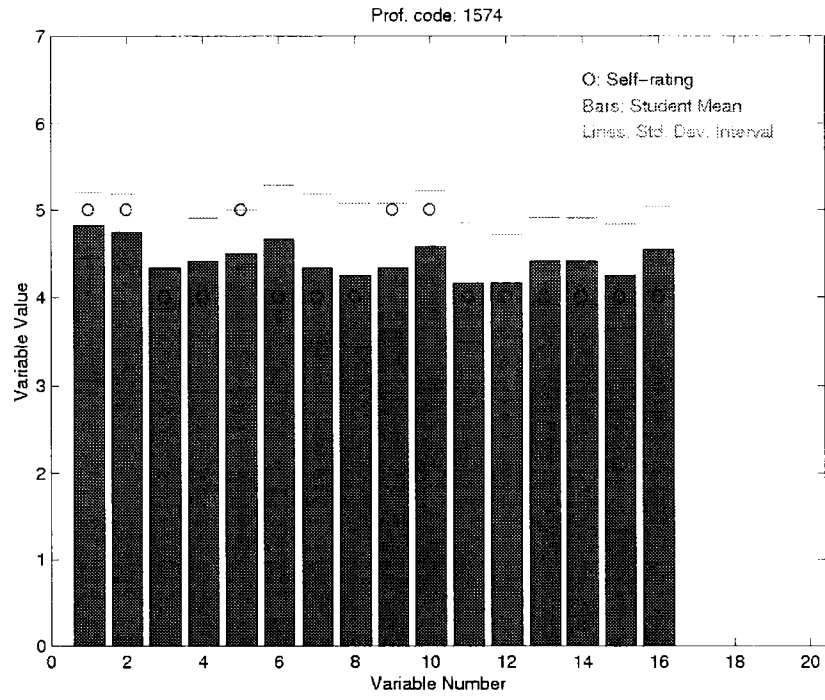


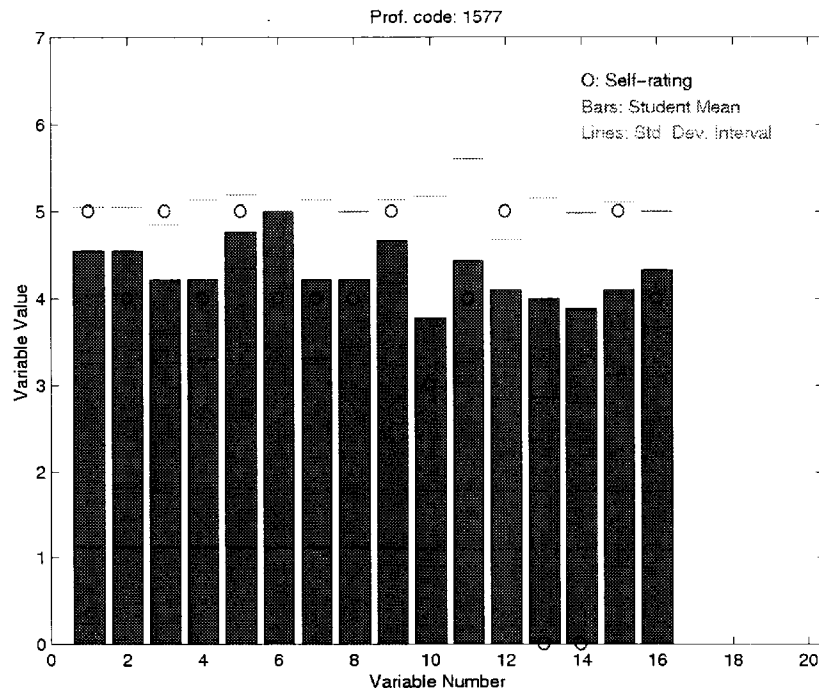
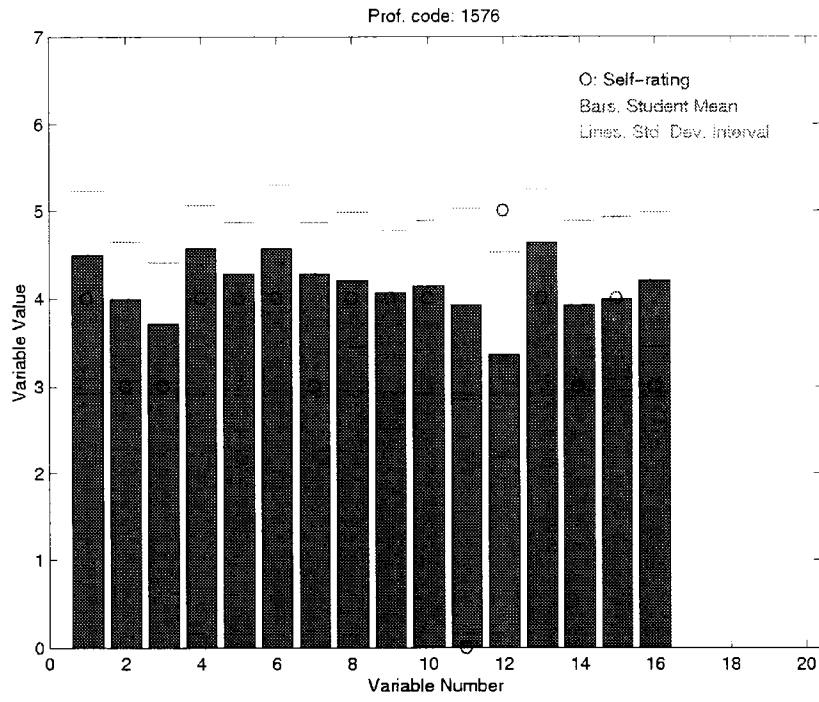




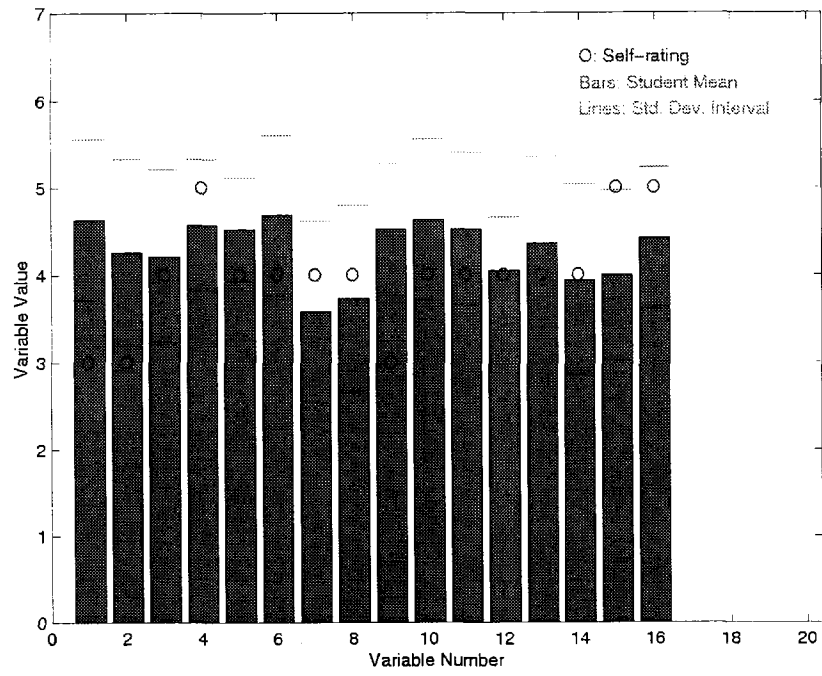








Prof. code: 1578



References

- Ad-hoc Committee on Student Evaluations-Memo to the Academic Committee, Ramapo College of New Jersey (2001). *Student evaluations and suggestions for reform*. Retrieved August 10, 2004, from <http://orion/ramapo.edu/facassem/studentevalsreport7-02.html>
- American Dictionary: A Random House Dictionary (1984). Toronto, Canada: Random House, Inc.
- Anderson, L. W., & Burns, R. B. (1989). *Research in Classrooms. The study of teachers, teaching and instruction*. New York: Pergamon Press.
- Bain, G. (1982). *Evaluating teaching: Purposes, methods, and policies*. Seattle, WA: University of Washington, Committee on the Evaluation and Improvement of Teaching and Development and Faculty Development Board. (ERIC Document Reproduction Service No. 289341).
- Barnett, C. W., Matthews, H. W., & Jackson, R. A. (2003). A comparison between student ratings and faculty self-ratings of instructional effectiveness. *American Journal of Pharmaceutical Education*, 67 (4), 1-6.
- Bosshardt, W., & Watts, M. (2001). Comparing student and instructor evaluations of teaching. *Research in Economic Education*, 32 (1), 3-17.
- Brodie, D. A. (1998). Do students report that easy professors are excellent teachers? *The Canadian Journal of Higher Education*, 28 (1), 1-20.
- Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research* (Individual Development & Educational Assessment-IDEA Paper No. 20). Manhattan, KS: Kansas State University, Division of Continuing Education, Center for Faculty Evaluation & Development. Retrieved May 9, 2004 from http://www.idea.ksu.edu/papers/Idea_paper_20.pdf
- Center for Teaching and Learning. (1994, September). Student evaluation of teaching. University of North Carolina at Chapel Hill. *for your consideration...suggestions and reflections on teaching and learning*, 16. Retrieved August 10, 2004, from <http://ctl.unc.edu/fyc16.html>
- Center for Teaching and Learning. (1997, Fall). Using student evaluations to improve teaching. Stanford University, Stanford, CA. *Speaking of teaching*, 9, Retrieved May 9, 2000 from http://ctl.stanford.edu/Newsletter/student_evaluations.pdf

- Centra, J. A. (1973). Self-ratings of college teachers: A comparison with student ratings. *Journal of Educational Measurement*, 10(4), 287-295.
- Centra, J. A. (1996). Identifying exemplary teachers: Evidence from colleagues, administrators, and alumni. In Svinidki, M. D., & Menges, R. J. (Eds.), *Honoring Exemplary Teaching, New Directions for Teaching and Learning*, No. 65 (pp. 51-56). San Francisco, CA: Jossey-Bass.
- Cimikowski, L., & Cook, J. (1996). A model instructional computing course for preservice teachers. Bozeman, MT: Montana State University. (ERIC Document Reproduction No. ED 398882).
- Coburn, L. (1984). *Student evaluation of teacher performance*. (ERIC Document Reproduction Service No. ED 289887). Retrieved May 9, 2000 from <http://ericae.net/db/edo/ED289887.htm>
- Dooris, M. J. (1997). *An Analysis of the Penn State Student Rating of Teaching Effectiveness*. A Report Presented to the University Faculty Senate of the Pennsylvania State University. Retrieved August 10, 2004, from <http://www.psu.edu/president/pia/cqi/srte/analysis.html>
- Dulz, T., & Lyons, P. (2000). Student evaluations: Help or hindrance? *Journal of Business Education*, 1, (No. 038-Proceedings Issue).
- El-Hassan, K. (1995). Students' ratings of instruction: Generalizability of findings. *Studies in Educational Evaluation*, 21, 411-429.
- England, J., Hutchings, P., & McKeachie, W. J. (1996). *The Professional Evaluation of Teaching*. (American Council of Learned Societies Occasional Paper No. 33). Retrieved August 10, 2004, from <http://www.acls.org/op33.htm>
- Feldman, K. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? *Research in Higher Education*, 28(4), 291-344.
- Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current, and former students, colleagues, administrators and external (neutral) observers, *Research in Higher Education*, 30 (2), 137-194.
- Flinders Foundations of University Teaching (FFOUT), Flinders University, Adelaide, South Australia (2001). Retrieved August 10, 2004, from <http://www.flinders.edu.au/teach/evaluate/what.htm>

- Fries, C. J., & McNinch, R. J. (2003). Signed versus unsigned student evaluations of teaching: A comparison. *Teaching Sociology*, 31(3), 333-344.
- Gordon, P. A. (n.d). *Student evaluations of college instructors: An overview*. Unpublished manuscript. Valdosta State University, Georgia. Retrieved on February 26, 2004, from <http://chiron.valdosta.edu/whuitt/files/tcheval.html>
- Gould, C. (1991). *Converting faculty assessment into faculty development: The director of composition's responsibility to probationary faculty*. Boston, MA: Conference on College Composition and Communication. (ERIC Document Reproduction No. ED 331068).
- Gray, M., & Bergmann, B. R. (2003). Student teaching evaluations: Inaccurate, demeaning, misused. *Academe*, 89(5), 44-46.
- Griffin, B. W., & Pool, H. (1998). Monitoring and improving instructional practices (and are student evaluations valid?). *Journal of Research and Development in Education*, 32, 1-9.
- Haskell, R. E. (1998). *Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century* (Report No. EDO-TM-98-08). Biddeford, MD: University of New England (ERIC Document Reproduction Service No. ED 426114).
- Hiltner, A. A., & Loyland, M. O. (1998). The effectiveness of annual faculty evaluations: Accounting faculty perceptions. *Journal of Education for Business*, 73 (6), 370-375.
- Howard, G. S., Conway, C. G., & Maxwell, S. E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology*, 77(2), 187-196.
- Howell, A. J., & Symbaluk, D. G. (2001). Published student ratings of instruction: Revealing and reconciling the views of students and faculty. *Journal of Educational Psychology*, 93(4), 790-796.
- Kaufman, B. J. (1981). *Departmental differences in student perceptions of 'ideal' teaching*. (ERIC Document Reproduction Service No. ED 212251).
- Lawall, M. L. (1998). *Students rating teaching: How student feedback can inform your teaching*. Retrieved May 9, 2000, from page 4 at http://umanitoba.ca/academic_support/uts/publications/seeq.pdf

- Leamon, M.H., Servis, M. E., Canning, R. D., & Searles, R.C. (1999). A comparison of student evaluations and faculty peer evaluations of faculty lectures. *Academic Medicine*, 74(10), S22-S24.
- Marincovich, M. (1998). *Ending the Disconnect Between the Student Evaluation of Teaching and the Improvement of Teaching: A Faculty Developer's Plea*. Stanford, CA: Stanford University School of Education. National Center for Postsecondary Improvement. (ERIC Document Reproduction Service No. ED 428590).
- 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., Hau, K. T., Chung, C. M., & Siu, T. L. (1997). Students' evaluations of university teaching: Chinese version of the students' evaluations of education quality instrument. *Journal of Educational Psychology*, 89(3), 568-572.
- Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). 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. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal*, 30(1), 217-251.
- Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist*, 52(1), 1187-1197.
- Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, 92(1), 202-228.
- Mason, K. H., Edwards, R. R., & Roach, D. W. (2002). Student evaluation of course instructors: A measure of teaching effectiveness or of something else. *Journal of Business Administration Online*, 1(2). Retrieved February 26, 2004, from http://jbao.atu.edu/mason_edwards_roach.htm

- Miller, G. (2003). Teacher evaluations often reflect attractiveness, study finds. State College, PA, Centre Daily Times (Knight Ridder/Tribune Information Services). Retrieved on August 10, 2004 from <http://www.highbeam.com/library/doc0.asp?DOCID=1G1:119395037&num=83&ctr1Info=Round....>
- Miller, J. L., Dzindolet, M. T., Weinstein, L., Xie, X. L., & Stones, C. R. (2001). Faculty and students' views of teaching effectiveness in the United States, China, and South Africa. *Teaching of Psychology*, 28 (2), 138-142.
- Moses, I. (1986). Self and student evaluation of academic staff. *Assessment and Evaluation in Higher Education*, 11(1), 76-86.
- Northwestern University, Chicago-Evanston, IL (1999), Searle Center for Teaching Excellence. *Teaching and Learning Issues. Student ratings and the evaluation of teaching: A White Paper*. Retrieved May 9, 2000, from <http://president.scfte.northwestern.edu/white.htm>
- Office of Institutional Research, Suffolk County Community College, Long Island, NY. (n.d.). *Student ratings of instruction-Report and recommendation*. Retrieved from August 10, 2004, from <http://instsrv.sunysuffolk.edu/strate.htm>
- Olp, M., Watson, K., & Valek, M. (1991). *Appraisal of faculty: Encouragement and improvement in the classroom*. Yuma, AZ: Arizona Western College (ERIC Document Reproduction Service No. ED 336159).
- Palmer, P. (1990, January). Good teaching: A matter of living the mystery. *Change*, 22(1), 10-16.
- Recker, M. M., & Greenwood, J. (n.d.). *An interactive student evaluation system*. Victoria University of Wellington, New Zealand. Retrieved August 10, 2004, from Utah State University, Instructional Technology Department Web site: <http://it.usu.edu/~mimi/papers/www.html>
- Reid, D. J., & Johnston, M. (1999). Improving teaching in higher education: Student and teacher perspectives. *Educational Studies*, 25(3), 269-281.
- Riger, S. (1993). *Gender and Teaching Evaluation*. University of Illinois, Chicago, Women's Studies Program. Retrieved August 10, 2004 from http://research.umbc.edu/~korenman/wmst/teaching_eval.html

- Robinson R., Arney, N., Munn, P., & MacDonald, C. (1990). *Perception of Effective Teaching Methods in Computer Studies* (SCRE Project Report). Edinburgh, Scotland: Scottish Council for Research in Education (ERIC Document Reproduction Service No. ED 324404)
- Ruskai, B. (1997). Evaluating student evaluations. *Notices of the American Mathematical Society*, 44 (3), 308.
- Sagen, H. B. (1974). Student, faculty, and department chairmen ratings of instructors: Who agrees with whom? *Research in Higher Education*, 2, 265-272.
- Schmelkin, L. P., Spencer, K. J., & Gellman, E. S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education*, 38(5), 575-592.
- Schwarz, J. (1997, December 4). Student evaluations don't get a passing grade: Easy-grading professors get too-high marks, new UW study shows. *News and Information from the University of Washington/uwnews.org*. Retrieved August 10, 2004, from <http://www.washington.edu/newroom/news/k120497.html>
- Scriven, M. (1995). *Student ratings offer useful input to teacher evaluations*. Washington, D.C.: The Catholic University of America, (ERIC Document Reproduction No. ED 398240).
- Seldin, P. (1993, July 21). The use and abuse of student ratings of professors. *The Chronicle of Higher Education*. A40.
- Senior, B. (1999). Student teaching evaluations: Options and concerns. In *Proceedings of the 35th Annual Conference of Associated Schools of Construction (ASC) at California Polytechnic State University-San Luis Obispo, California, April 1999* (pp.251-260). Ames, Iowa: Office of Editorial Services, College of Engineering.
- Siegel, M. E. (1985). *The challenges of improving the teaching-learning process in computer studies*. Adelphi, MD: The University of Maryland's University College, (ERIC Document Reproduction No. ED 258648).
- Siegel, M. E., & Johnstone, S. M. (1985). *The challenges of improving the teaching- learning process in computer studies*. Adelphi, MD: The University of Maryland's University College.

- StatSoft, Inc. (2004). Nonparametric statistics (chap.). In *Electronic Statistics Textbook*. Retrieved November 16, 2004, from <http://www.statsoft.com/textbook/stnonpar.html>
- Stevens, J. J. (1987). Using student ratings to improve instruction. In Aleamoni, L. M. (Ed.), *Techniques for Evaluating and Improving Instruction, New Directions for Teaching and Learning, No. 31* (pp. 33-38). San Francisco, CA: Jossey-Bass.
- Student Evaluations: A Critical Review. (n.d.). Retrieved August 10, 2004, from <http://home.sprynet.com/~owl1/sef.htm>
- Tang, T.L. (1997). Teaching evaluation at a public institution of higher education: Factors related to the overall teaching effectiveness. *Public Personnel Management, 26*(3), 379-388.
- Tripe, R. L. K. (1990). *Problem solving and writing: A teaching/learning model for computer studies*. Ontario, Canada: Mohawk College, Computer Science Department. (ERIC Document No. ED 324039).
- University of Michigan, Ann Arbor, MI (2004), Center for Research on Learning and Teaching. *Guidelines for evaluating teaching*. Retrieved August 10, 2004, from <http://www.crlt.umich.edu/crlttext/guidelinestext.html>
- Watchel, H. K. (1998). Student evaluation of college teaching effectiveness; a brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191-211.
- White, L. J. (1995). Efforts by departments of economics to assess teaching Effectiveness: Results of an informal survey. *Journal of Economic Education, 26*(1), 81-85.
- Wilson, R. (1998, January 16). New research casts doubt on value of student evaluations of professors. *The Chronicle of Higher Education*, pp.1-5.
- Younglich, A. (1955). Study on correlations between college teachers' and students' concepts of "ideal-student" and "ideal-teacher." *Journal of Educational Research, 49*, 59-64.

CURRICULUM VITAE

LAURIE SCHWARTZ NAPARSTEK
34 Vassar Street, Worcester, MA 01602
(508) 754-7631

EXPERIENCE

4/02-Present **Human Resource Assistant**, Public Sector Partners, Inc./UMass Medical School
Worcester, MA

- Assistant to the Human Resources Director; oversee all employee related materials including confidential salary information, benefits materials, etc.
- Responsible for reference calls, scheduling interviews, maintaining employee database, interfacing with employees and troubleshooting as needed.

1/98-11/01 **Administrative Assistant**, The Kenneth B. Schwartz Center/MA General Hospital
Boston, MA

- One of three employees staffing the center named in memory of my late brother who passed away of lung cancer in 1995 at age 40 at Massachusetts General Hospital.
- Responsible for providing administrative support to executive director, administrative director, board members and committees.

9/90-2/97 **Administrative Assistant**, B.U. Metropolitan College/Computer Science Dept.
Boston, MA

- Provided administrative direction for the largest academic department in Metropolitan College, an extension school serving working adults at four campuses in and around the Boston area.
- Acted as the liaison between the college, seven full-time faculty members and over 50 adjunct professors that involved recruitment, course assignments, the long-range design of class schedules/catalogues and mediation/trouble shooting.
- Responsibilities included the hiring and supervision of office staff, teaching assistants and graduate assistants.

6/88-6/89 **Human Resource Representative**, Apollo Computer, Inc./Human Resource Dept.
Chelmsford, MA

- Updated, expanded and redesigned the Human Resource Policy and Procedure manual for Apollo's 3200 U.S.A. based employees.
- Member of a task force that planned the implementation of an Employee Assistance Program.

Intern

- Solely responsible for recreating and implementing a company-wide New Employee Orientation Program. Used innovative approaches to introduce new employees to the history, culture, philosophy, products and benefits of the company.
- Recipient of "Apollo Achievement Award" upon completion (9/87-4/88).

3/85-1/87 **Field Coordinator**, Medical Register Inc. Boston, MA

8/84-9/85 **Registrar's Assistant**, Cambridge Center for Adult Education Cambridge MA

1/83-1/84 **Aerobics Director**, The Exercise Company Brookline, MA

5/81-7/82 **Activities Director**, Franklin County Mental Health Center Greenfield, MA

EDUCATION

9/89- Present	Boston University Doctor of Education candidate in Developmental Studies, expected May, 2005	Boston, MA
6/89	Harvard University Master of Education in Human Development and Psychology	Cambridge, MA
6/88	Suffolk University Master of Science in Human Resource Development	Boston, MA
12/86	Bentley College Human Resources Management Certificate	Waltham, MA
5/81	University of Massachusetts Bachelor of Science in Human Development Requirements completed for a major in Community Services Minor: Psychology Graduated cum laude	Amherst, MA

HONORS

Pi Lambda Theta (National Honor and Professional Association in Education)
Kappa Omicron Nu (National Scholastic Honorary Society in Community Services)

BIRTH YEAR 1959