The Effect of an Educational Intervention on Faculty and Administrator Knowledge and Attitude to Student Course Evaluations

Michael A. Rossi Jr.
Seton Hall University

Follow this and additional works at: https://scholarship.shu.edu/dissertations

Part of the Adult and Continuing Education Administration Commons, and the Other Education Commons

Recommended Citation
Rossi Jr., Michael A., "The Effect of an Educational Intervention on Faculty and Administrator Knowledge and Attitude to Student Course Evaluations" (2002). Seton Hall University Dissertations and Theses (ETDs). 1549.
https://scholarship.shu.edu/dissertations/1549
THE EFFECT OF AN EDUCATIONAL INTERVENTION ON FACULTY AND ADMINISTRATOR KNOWLEDGE AND ATTITUDE TO STUDENT COURSE EVALUATIONS

BY

MICHAEL A. ROSSI, JR.

Dissertation Committee

Martin Finkelstein, Ph.D., Mentor
Joseph Stetar, Ph.D.
Shouping Hu, Ph.D.

Submitted in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy Seton Hall University 2002
ABSTRACT

This study tested the impact of an educational intervention on faculty and administrator knowledge and attitude to student course evaluations. Student course evaluations are one of the most controversial and polarized topics in higher education. Based on the weight of research evidence that has established the reliability and validity of student course evaluations, virtually every college and university employs them as a primary tool to measure teaching effectiveness. However, faculty often dismiss them as unfair, unreliable and susceptible to biasing influences. The source of that contradiction, lack of knowledge of obtaining, interpreting and communicating data related to student course evaluations, was a focus of this study. The research question asked: to what extent does level of user knowledge affect corresponding attitudes toward student course evaluations? Subsidiary questions asked to what extent can both knowledge and attitude be changed through educational interventions and how intensive an intervention was required to change knowledge or attitude? The study employed a quasi-experimental design to test the effect of educational interventions on faculty and administrator knowledge and attitude to student course evaluations. Ninety-three participants (from twenty-six New Jersey colleges and universities), in three separate groups, comprised the sample of this study. Two distinct interventions were tested in the research. All three groups completed a pre and post survey on the fundamental knowledge variables of student course evaluations, and also responded to attitude questions on student course evaluations. The Workshop Group completed the survey prior to, and at the conclusion of a workshop on student course evaluations. The Reading Group completed the survey prior to, and at the conclusion of reading three articles on student course evaluations. The Control Group completed the pre and post survey without an intervention. The hypotheses were that subjects with higher knowledge will demonstrate significantly more favorable attitudes toward student course evaluations, that administrators would demonstrate significantly more knowledge and significantly more favorable attitudes toward student course evaluations, and that those subjects afforded more intensive educational experiences will demonstrate significantly higher knowledge and favorability toward student course evaluations. Study results indicate that a significant correlation
between knowledge and attitude was demonstrated on both the pre and post survey ($p < .001$). Administrators did not demonstrate higher knowledge or more favorable attitudes, on either the pre or post survey. The impact of the educational interventions did indeed improve both attitude and knowledge, for both the Workshop and Reading Groups ($p < .01$), as compared to the Control Group. The change within the Workshop and Reading Groups in knowledge ($p < .01$), was significant, as was the attitude change within the Workshop ($p < .05$) and Reading Group ($p < .01$). The Control Group did not demonstrate any significant knowledge or attitude change. Recommendations for further research include analyzing the three knowledge domains separately, replicating the study to account for institutional and discipline differences, and conducting a follow-up survey of the Workshop and Reading Groups to measure retention of knowledge and attitude.
ACKNOWLEDGEMENTS

A great many people have sacrificed their time and energy to help me complete this process. A few of them are mentioned here:

Debora Rossi, my wife, for sharing all the ups and downs of graduate school.

My parents, Michael and Mary Rossi, for always reminding me of what is really important.

My children, Kaitlyn, Michael III and Lillian Rose, for being what is important.

My mentor, Dr. Martin Finkelstein; his “best” kept me moving “onwards.”

My sister Celeste and her husband Gordon Holovacs, for their outstanding “technical assistance” and moral support.

Dr. Joseph Stetar, for teaching me so much about higher education.

Dr. Shouping Hu, for his statistical advice on my dissertation.

Dr. Anthony Colella, for being a great advisor.

Mrs. Carol Rawnsley, for literally getting me through Seton Hall.

My brothers, Joseph and Anthony; my sister Roselyn, for always being there for me.

Bernard and Karen Ivin, for their magnificent help when we needed it most.

My in-laws, James and Margaret Goldworthy, for their babysitting while I worked.

My colleagues at Centenary College, who followed my path from a non-matriculated student to Ph.D.

Mr. Dale Hauser, for his steady navigation through a long journey.

Mr. John Cullinane, for his positive reinforcement of all situations.

The 2001-02 faculty of Hope Township Elementary School, for their collegiality and warm hearts.
DEDICATION

This dissertation is dedicated to my father,

Michael A. Rossi;

the best teacher I've ever had,

both in and out of the classroom.
# TABLE OF CONTENTS

**ACKNOWLEDGEMENTS** ........................................................................................................ iii

**DEDICATION** .................................................................................................................... iv

**LIST OF TABLES** .................................................................................................................. vii

**LIST OF FIGURES** ................................................................................................................... viii

I. **STATEMENT OF THE PROBLEM** .................................................................................. 1

   Background .......................................................................................................................... 1
   Definition of the Problem ...................................................................................................... 5
   Purpose of the Study ............................................................................................................. 10
   Research Question ............................................................................................................... 11
   Subsidiary Questions ............................................................................................................ 11
   Study Summary .................................................................................................................... 11

II. **LITERATURE REVIEW** .................................................................................................. 13

   Background of Student Course Evaluations ......................................................................... 13
   The Growth and Utility of Student Course Evaluations ......................................................... 13
   Defining the Criteria of Effective Teaching ........................................................................... 16
   Typical Faculty Concerns About Student Course Evaluations ............................................. 18
   Establishing the Reliability and Validity of Student Course Evaluations: Destructing .......... 23
   Multi-Section Validity Studies ............................................................................................... 23
   Multi-Method Studies .......................................................................................................... 24
   Bias Studies ........................................................................................................................... 25
   Meta-Analytic Studies .......................................................................................................... 26
   Student Course Evaluations: Designing a Best Practice Approach ....................................... 28
   Knowledge of Obtaining Data ............................................................................................... 29
   Knowledge of Interpreting Data ........................................................................................... 31
   Knowledge of Communicating Data ..................................................................................... 32
   Summary of Literature Review ............................................................................................. 33

III. **METHODOLOGY** ........................................................................................................... 35

   Research Design ................................................................................................................... 35
   Study Participants ................................................................................................................ 36
   Population ............................................................................................................................. 36
   Sample ................................................................................................................................... 36
   Educational Interventions ..................................................................................................... 44
   Hypotheses ........................................................................................................................... 45
   Data Collection ..................................................................................................................... 46
Conceptualizing the Variables ........................................ 46
Operationalizing the Variables ................................... 47
Instrument ................................................................. 50
Data Collection Procedures ......................................... 51
Data Analysis .............................................................. 53
Limitations ................................................................. 57

IV. DATA ANALYSIS AND FINDINGS ........................................ 59
Sample Characteristics .................................................. 59
Quantitative Findings .................................................... 63
Hypothesis 1 and 2 ....................................................... 63
Hypothesis 3 ............................................................... 65
Hypothesis 4 ............................................................... 67
Hypothesis 5 and 6 ....................................................... 70
Further Examination of Hypothesis 5 and 6 ..................... 73
Summary of Data Analysis ............................................. 75

V. DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS .......... 79
Introduction .................................................................. 79
Discussion of Results ..................................................... 82
Conclusions ................................................................. 87
Recommendations for Further Research and Practice ............ 91

REFERENCES ................................................................ 96

APPENDICES ................................................................ 109
A. Workshop Group Recruitment Letter .......................... 110
B. Colloquium Registration Form .................................... 113
C. Reading Group Recruitment Letter ............................. 115
D. Reading Group Instructions ........................................ 117
E. Theall & Franklin (1990) ............................................ 119
F. Franklin & Theall (1990) ............................................. 137
G. Theall & Franklin (1991) ............................................ 157
H. Recruitment Letter for Control Group ......................... 175
I. Franklin & Theall (1989) survey .................................. 177
J. Expert Group for 1989 survey ..................................... 188
K. Instrument used in this Study ..................................... 190
L. Numeric Equivalents of 6-25 on Instrument .................. 194
LIST OF TABLES

1. Sample Characteristics: Rank .................................................................59
2. Sample Characteristics: Experience .....................................................60
3. Sample Characteristics: Current Duties ................................................60
4. Sample Characteristics: Participation ...................................................60
5. Sample Characteristics: Creating or Selecting Forms ............................61
6. The Relationship Between Knowledge and Attitude on the Pré Survey .......64
7. The Relationship Between Knowledge and Attitude on the Post Survey .....65
8. Knowledge and Attitude Differences (Administrator verses Faculty) ANOVA .................................................................66
9. Impact of Educational Interventions on Knowledge and Attitude (Descriptives) .................................................................68
10. Impact of Educational Interventions: ANOVA (Pre Survey) ....................69
11. Impact of Educational Interventions: ANOVA (Post Survey) ..................69
12. Impact of Educational Interventions: Least Significant Difference ..........70
13. Knowledge and Attitude Change Within Groups from Pre to Post Survey: Workshop Group .................................................................72
14. Knowledge and Attitude Change Within Groups from Pre to Post Survey: Reading Group .................................................................72
15. Knowledge and Attitude Change Within Groups from Pre to Post Survey: Control Group .................................................................73
16. Knowledge and Attitude Change Between Groups: ANOVA ....................74
17. Least Significant Difference Test of Change .........................................75
18. Summary of Data Analysis ..................................................................77
LIST OF FIGURES

1. Faculty Rank ........................................................................................................38
2. Years of Experience ...............................................................................................38
3. Current Duties .........................................................................................................39
4. Levels of Participation in Personnel Decisions ...................................................39
5. Levels of Participation in Creating or Selecting Forms ......................................40
6. Participation by Group ..........................................................................................42
7. Colleges and Universities Represented ...............................................................42
CHAPTER I

STATEMENT OF THE PROBLEM

Background

The end of semester student course evaluation of teachers has become a standard component of evaluating teaching effectiveness and student learning in American higher education. A 1993 study of 600 liberal arts colleges revealed that 86% of them were using student course evaluations (Seldin, 1993). A review of the literature suggests that virtually every college in America employs student course evaluations as a means of evaluating college courses, and many feel student course evaluations are the primary, if not sole measure of teaching performance and student learning (Magner, 1994). Moreover, information gleaned from these questionnaires often play a major role in determining faculty personnel decisions (Cashin & Downey, 1992).

By virtue of their dominance in the assessment of college teaching, student course evaluations has been the subject of many research studies. The focus of almost all this research was to determine if student course evaluations were reliable and valid instruments to judge teaching and learning. Woven into these studies were extraneous variables that may impact student course evaluation results. They include class size and level, students' grades, teacher personality and background, among
others. Academics have lamented the potential for non-academic biases such as these to contaminate ratings.

Naftulin, Ware and Donnelly (1973) warned of the “Dr. Fox Effect,” where an actor posing as an instructor was rated highly effective by students despite lectures devoid of meaning. Divoky and Rothermel (1988) posited that student course evaluation results depend on course type (i.e. large lecture vs. small seminar), and as such outcomes should be adjusted to reflect those differences. In general, smaller classes (less than 15 people) are rated most highly, followed by those with 16-35 people, and those with over 100. The lowest scores tend to come from classes with 36-100 people (Theall, Franklin, 1991). Concomitantly, class size has been established by several research studies, as affecting ratings outcomes (Cranton & Smith, 1990; Murray, Rushton and Paunonen, 1990). Holmes (1972), Kolevzon (1981), and Cassuto (1998) to name but a few have cited the issue of grading leniency (better grades mean better ratings). Thus, there are arguably more than just “pockets” of dissent regarding perceived validity and reliability of student course evaluations to assess teaching and learning.

There are, however, at the same time, volumes of research studies, accompanied by data that affirm student course evaluations as a reliable measure of stable dimensions of teaching effectiveness. Researchers have endorsed the validity and reliability of student course evaluations by correlating student ratings of particular teaching traits, observations by external evaluators, self-evaluations, and responses from former students. Some of the major contributors to this conception include Marsh (1987, 1984, 1983); Miller (1987); and Centra (1993), among many
others. Moreover, Howard and Maxwell (1980), Abrami, d' Appollonia, and Cohen (1980), and Marsh and Roche (2000) found little or no connection between student course evaluation outcomes and what they claim as the “grading bias myth.”

Centra and Creech (1976) studied responses from 300,000 students, in 16,000 classes, at over 100 colleges and found “insignificant relationships” between student ratings and age, sex, grade point average or academic ability. Feldman (1976a, 1976b, 1978, 1989a, 1989b), synthesizing numerous studies found little correlation between faculty age, background, or gender and student evaluations. Additionally, Feldman’s research posits that student ratings of teachers are most accurate on the following dimensions of teaching: preparation and organization, clarity, perceived impact of instruction, stimulation of interest, enthusiasm, and other global items such as “overall effectiveness.” They are especially accurate when the student is asked to gauge only his or her own individual experience through low-inference questions.

A low inference question can be defined as one that asks each student to rate the professor on a given teaching dimension as it impacted only his or her personal experience. For example, a question that asks: "the professor stimulated my interest to learn in this course," is low inference because it only asks the student to comment on what he or she experienced. Conversely, a question that asks: "the professor stimulated interest in this course," is asking the student to comment on each learner’s experience. That type of question is high inference and a less reliable measure.

Thus, the weight of empirical evidence suggests that, through low inference questions there are a handful of teaching dimensions that student course evaluations can measure validly and reliably. First, low inference questions on clarity can be
measured quite accurately through student course evaluation reports. Students can provide correct evaluations of clarity by responding to questions that ask if the professor gave frequent examples, used graphs or diagrams, suggested ways to retain knowledge components, among others. Second, low inference questions on stimulation of interest can be accurately gauged by asking students to react to questions that ask how much time was spent outside of class studying, how valuable time spent studying was to them, the degree to which assignments helped students achieve course goals, among others. Students can provide accurate evaluations of perceived impact of instruction by asking them questions about how much they have learned from the course, how the course compares to others in terms of overall learning, and whether they feel the instruction was adequate to their personal learning needs, among others. Students can also accurately respond to questions on preparation and organization. Students can provide accurate evaluations of overall effectiveness of both the course and the professor by asking students to compare the particular course in question with all others they have taken and rating the professor against all other professors they have taken. These two items (overall effectiveness of course and instructor) even have construct validity when the rest of the student course evaluation form does not. Moreover, balancing them against peer evaluations and other relevant comparisons has validated these two components.

Despite the weight of the data endorsing student course evaluations as a valid and reliable measure of teaching and learning, opinion pieces that decry their use continue to appear in large numbers. Some speculate that the variety of questionnaire structures and variety of responses within the same class makes feedback seem
meaningless (Gold, 1997). Others believe a professor who uses "all five grades" will suffer from bad Student course evaluation ratings (Gose, 1997). Haskell (1997a, 1997b, 1997c) claim student course evaluation use is a threat to academic freedom, and Perillo (2000) tries to justify not even reading course evaluations. Negative sentiment continues to increase even as empirical evidence of reliability and validity also increases.

Definition of the Problem

Heavy emphasis within the literature essentially bears upon two premises. One is comprised of large research studies supporting the validity and reliability of student course evaluations. The other is comprised of smaller studies or anecdotal evidence, which stress individual, specific examples of how and why they do not work as an evaluation of teaching and learning. The preponderance of research evidence supports the former but faculty opinion often focuses on the latter.

What accounts for this contradiction is the focus of this study. Franklin and Theall (1989, 1990, 1999) suggest that the source of this contradiction lies not in the instrument used, but rather in a lack of knowledge of how student course evaluations should be used, in a lack of statistical skills and in a lack of knowledge in how to communicate results to stakeholders. According to their research, it is not the psychometric properties of the instruments that reduce the effectiveness of student course evaluations but rather bad practice by users. Individuals responsible for developing, administering, interpreting and communicating ratings outcomes need to be knowledgeable in the areas of obtaining, interpreting and communicating student course evaluation data.
It seems quite possible that it is bad practice by users that cause academics to bitterly complain about student course evaluations. It also seems quite plausible that the oft-negative publicity surrounding student course evaluations results from a lack of knowledge about the variables that make them valid as measures of teaching and learning. Both of these possibilities negate the developmental potential of student course evaluations as a vehicle for improving teaching and learning. It also appears that the testimonies of how student course evaluations have negatively impacted faculty could arguably be the result of bad practice due to lack of knowledge.

This lack of knowledge centers on how aware users are of the variables that influence best practice for the use of student course evaluations. A user can be described as anyone who is impacted or influenced by student course evaluation results. They include students, whose academic choices can be influenced by published student course evaluation results. They include faculty, for whom student course evaluation results may require anything from changing how they teach to looking for new employment. They also include anyone responsible for making personnel decisions that include data from student course evaluation outcomes. The last two are the focus of this study.

Based on Franklin and Theall (1990), there are three "safeguards and standards for practice" that help users determine the most appropriate student course evaluation format and to make the best sense out of their results. The first consideration is the method selected for obtaining data. The structural validity of any student course evaluation form will depend upon how familiar its creators are with the literature surrounding student course evaluation questionnaire development. In
addition to the method, obtaining data also includes selecting the right instruments to process and manage the data. The second consideration involves user ability to analyze data resulting from ratings. The literature informs us that how results are interpreted is just as important as selecting the right focus questions. The third consideration is the process used to communicate data to users. This also is a critical stage because as the literature reveals student course evaluation results must be put into appropriate contexts for users to make sense of them. Moreover, this stage includes any follow-up or developmental approaches implemented; these measures would be in reaction to student course evaluation results and designed to improve teaching and increase learning.

User knowledge of each of the three variable sets influences the overall effectiveness of any student course evaluation design and practice. Thus, it is hypothesized that much of the acrimony surrounding student course evaluations is a result of inappropriate use related to a lack of knowledge. Student course evaluations are a valid and reliable means to gauge teaching effectiveness and student self-perceived learning. However, their implementation across dozens of disciplines, departments, and thousands of courses is a very complex practice. Proper use requires systematic planning, the ability to analyze data and place it within a particular context, and a clear and formative method to communicate results. One message from the literature is that too many users are not knowledgeable enough about student course evaluation variables to devise an evaluation system that provides valid, beneficial information. A close look at the attributes of the three primary knowledge variables underscores the complexity of student course evaluations. In
terms of selecting the questionnaire items to obtain data, users must have knowledge of the type of questions that have been found to be most reliable and valid. Much evidence points to the validity of global items as a measure of teaching and learning (Centra, 1993). Questionnaires should focus on "low-inference" items because they are the best measure of faculty effectiveness (Murray, 1983). In addition, questions need to be clear and cover only one topic at a time (Theall & Franklin, 1991). For instance, students can respond with accuracy to items related to their own experience, such as questions that ask if the pace of the course was one they could follow. Students can respond with accuracy to questions about clarity of expectations and how a particular professor rates in comparison to others.

Unfortunately, Ory and Wieties (1991) found that many questions on student course evaluation forms are of doubtful value. Many questionnaires for example, ask for student responses to how "knowledgeable" professors are. This question in particular, combined with others such as grading policies tends to obfuscate overall results. Compounding the situation is the often-problematic wording of questions, changing them from low to high-inference items. For example, asking students to respond to ("course expectations were clear to me") is low-inference while asking them to respond to ("course expectations were clear to students") is high inference.

In terms of interpreting data users need to be familiar with the literature on best practice for interpreting student course evaluation reports. Awareness of sample size and acceptable response ratio is one example (such as in a class of 30 a minimum of seventy-five percent response is needed). Another example is the awareness and ability of the interpreter to place an individual's score into a percentile group.
equivalent (or vice versa). In some situations professors' percentile rank may be higher against campus means than against department means. The opposite might also be true. How to interpret a t or other standardized score as it relates to an individual, department, or campus norm is critical to the validity and reliability of any student course evaluation system. Accordingly, Franklin and Theall (1989) found that "users may not know all they need to know about using student ratings" (p. 24).

In terms of communicating data to clients, users must confirm the ability of all stakeholders on the receiving end of data to make sense of it. In addition to validating the construct of the instrument and results, communicators must validate that all constituents within the student course evaluation process understand the meaning of the results. Communication of results must also include a view of them over time, taking into account teaching loads over time, course levels, and particular areas of student course evaluation ratings that may have changed (improved) over time.

User awareness of student course evaluation knowledge variables has received scant attention in the literature. The vast majority of articles focus on validity and reliability. Only recently have some scholars addressed the issues of knowledge and awareness of the fundamental considerations in student course evaluation use as it relates to best practice. Users must be capable of verifying appropriate sampling size, determining if results are significant or not, and ultimately be able to point to strategies for improvement. In comparison to the vast amount of literature on reliability and validity, very little has been written about how this can be accomplished. More importantly, two things are apparent from the severity of
complaints about student course evaluations, which belie research results. First, it is evident that too many users are not familiar enough with the important variables that are required to ensure well-founded designs, interpretations, and communications of student course evaluation results. Second, the complexity of any ratings system requires training for all those involved. Simply creating a questionnaire is not enough, a ratings system that is used in personnel decisions must be accompanied by an increase in user knowledge of the variables that impact good use.

Purpose of the Study

The purpose of this study is to test the relationship between level of knowledge of the three primary student course evaluation variables and attitude toward their use. Based on existing literature reviews, it is hypothesized that negative views regarding student course evaluation use are related to lack of knowledge. Further, it believed that there is a relationship between a user's knowledge level and his or her corresponding attitude toward ratings issues. With that in mind, this study is a follow-up to the 1989 study conducted by Franklin and Theall. The study seeks to confirm the relationship between knowledge and attitude reported in their study.

More importantly, the study is an investigation of the impact of educational interventions on knowledge. The study seeks to determine if educational interventions can increase knowledge and if subjects who demonstrate an increase in knowledge then also demonstrate a more favorable attitude toward student course evaluations. As subjects become more knowledgeable, they will demonstrate a more favorable attitude. Conversely, at lower levels of educational intervention (including no intervention) subjects will respond with less or no increase in knowledge and no
change in attitude. In sum, it is the purpose of this study to test whether an educational intervention raises knowledge levels, and whether it has a corresponding affect on attitudinal favorability. As a result of the research of this study it is hoped that a recommendation can be made regarding student course evaluation best practice.

Research Question

1. To what extent does level of user knowledge of student course evaluation variables affect corresponding attitudes toward student course evaluation issues?

Subsidiary Questions

1. To what extent can such attitudes be changed by educational interventions designed to increase knowledge?

2. How intensive an educational intervention is required to train users in the fundamental components of student course evaluations to significantly increase knowledge or improve attitudes?

Study Summary

Student course evaluations are a practical mechanism to use as a gauge for evaluating teaching effectiveness and as a basis for faculty personnel decisions. Their existence gives students a voice, teachers' feedback on their performance, and administrators a tangible glimpse into what is happening in college classrooms. A review of the literature finds strong empirical evidence to support their ability to reliably measure global items related to teaching and learning. It is hypothesized that opposition to their use has resulted from a lack of knowledge; of what they measure best, of what the data results really mean (and the requisite knowledge to make those determinations), and resulting from the inability of many users to effectively
communicate outcomes. It is further hypothesized that educational interventions of increasing intensity levels may significantly increase knowledge and thereby enhance attitudes.

Despite the breadth of studies verifying their reliability and validity, lack of knowledge associated with student course evaluation use and ratings systems by users of all kinds (those being judged and those doing the judging) has led to misuse, suspicion, and hostility. The variables of obtaining, analyzing, and communicating data include a variety of information with which users need to be familiar. Taking into account the depth of the reliability/validity studies, combined with the continued accounts of misuse and general lack of acceptance among faculty, it is hypothesized that there are not enough users with a complete understanding of the knowledge and skill base associated with student course evaluation ratings systems. It is also hypothesized that an increase in knowledge will result in an increase in competent use and favorable attitudes.
CHAPTER II

LITERATURE REVIEW

Background of Student Course Evaluations

The Growth and Utility of Student Course Evaluations

In American higher education quality assessment of teaching and learning has long centered on the tenuous balance between accountability and autonomy. While professors often feel their classrooms are their personal domains, reform-minded individuals stress the need for external assessment of teaching and learning. Since a Nation at Risk in 1983 (National Commission on Excellence in Education, 1983) and continuing with Goals 2000 (US Department of Education, 1990), higher education has experienced a steady growth of external pressures that reflect a "consumer" climate. Outcomes assessments and performance indicators, such as graduation rates, number and types of services provided and publication of national qualifying exams (such as the number of students passing the Praxis Teacher's Exam), are all indicators of heightened accountability through qualitative as well as quantitative measures.

Measuring teaching effectiveness and student learning are now stable components of virtually every assessment and accountability process. Formative and summative assessment of a teacher's professional acumen takes several forms. However, the most commonly used assessment measure of classroom teaching is the end of semester student course evaluation. Close to 90% of all American college and
universities utilize student course evaluations to measure teaching and learning (Selhin, 1993). Additionally, colleges and universities often use student course evaluations as a primary component in faculty personnel decisions (Magner, 1994, Cashin & Downey, 1992).

Student course evaluations have several applications in college. Originally, student course evaluations were implemented to supervise teaching quality and to offer suggestions for teaching improvement (Guthrie, 1954). Today, ratings results from student course evaluations provide several functions. Administrators use student course evaluations to promote and grant tenure, hire adjuncts to full-time positions, decide on merit pay increases, assigning courses and more. Teachers use student course evaluations to improve teaching, as part of their own self-assessments and for portfolio purposes (Kulick, 2001).

The rationale behind the increased use of student course evaluations is essentially that students should be allowed to evaluate their own learning, and in fact, are the best suited of all higher education stakeholders to assess what they have learned and how well they were taught (McKeachie, 1969). Moreover, it should be an aim of higher education to inculcate in students “the ability to evaluate their own learning” (McKeachie, 1979, p.386). Further justification for the use of student course evaluations is their efficacy toward improving instruction. Aleamoni (1974) cites the connection between alumni and student ratings while Centra (1973) and Pambookian (1976) demonstrated that faculty could improve their teaching by comparing self-ratings to those of their students. Overall and Marsh (1977) found a correlation of .59 between student ratings at the end of the semester and the ratings of
those same students a year later. Overall and Marsh (1979) cited improved ratings through the use of mid-semester feedback. Overall and Marsh (1980) again tested the correlation between end of semester reports and alumni ratings. They asked 1,400 students to rate teachers of 100 courses; some students were rating instructors on courses they had taken three years prior. They found a .83 correlation between the original student course evaluations (at the end of the term) and student ratings as alumni. Thus, as the use of student course evaluations began to rise, research designed to validate their efficacy was also underway.

At the same time, the notion of students as consumers grew in its scope. Higher education became more responsive to the concerns and opinions of students. Giving students a voice in the educational process took many forms; using course evaluations to measure teaching effectiveness was a natural result. Various sources cite the use of student course evaluations as a major component of faculty personnel decisions (Magner, 1995; Marchese, 1997). Many researchers have examined (and added to) the literature on course evaluations as a means for teaching improvement and personnel decisions. Theall and Franklin (2001) state that by educating faculty and administrators, colleges and universities can design a ratings system that does both. Kulik (2001) points out that student ratings are so closely matched with other forms of faculty evaluation, they are useful as a component of all formative and summative evaluations. Additionally, Ory and Ryan (2001) suggest that any ratings system has the potential to be valid and reliable. It is the users that make a system successful.
Defining the Criteria of Effective Teaching

The utility of student course evaluations has risen steadily since the 1960s. With that rise, researchers have attempted to identify the primary components of good teaching. The rationale behind this approach was built on three major premises. The first was that there are fundamental aspects of "good teaching" that are stable across all teaching dimensions. That is to say, that regardless of the subject or level, good teaching can be defined by similar traits. The second was that good teaching would lead to better educational outcomes. The third was that those stable components of good teaching would also result in higher ratings on student course evaluations. So the foundation of "effective teaching" research tested connections between teaching, learning and the results of student course evaluations.

Several studies during the 1970s and 1980s sought to define effective teaching traits. Researchers have used various measures to identify effective teaching. Initially, studies sought to link achievement on exams as a preliminary connection between teaching and learning. While test scores are not the ideal method of evaluating teaching, early studies such as Frey (1973), McKeachie and Kulik (1975), and Kulik and McKeachie (1975) established positive correlations between student achievement and student ratings. These and other studies such as Sullivan and Skanes (1974), and Centra (1977) found correlations of .4 to .6 between student ratings and student achievement. These early studies provided two important considerations for researchers of student course evaluations. First, there seemed to be a connection between teaching effectiveness (provided through ratings results) and student achievement (provided through test results). Second, given the relationship
between teaching and achievement, it would be important to identify effective
teaching traits that were stable across all teaching dimensions. Moreover, it would be
important to determine which of those teaching traits could be accurately measured
by student course evaluations.

Researchers began to test aspects of effective teaching. A large body of
research emerged on university students’ opinions of college teaching (and specific
teachers and courses they had). Studies examined professors’ enthusiasm,
knowledge, clarity, fairness, availability, workload requirements, feedback and many
other facets of college teaching. Some studies examined student views on the
characteristics important to good teaching, the ideal teacher or the students’ best
teachers. Those studies tried to correlate how students ranked those characteristics
with overall evaluations of instructors made by students (through student course
evaluations). Other studies looked at specific teaching characteristics found on
student course evaluations (clarity, knowledge, etc.), and how closely associated
individual characteristics were with the global evaluation of the instructor (the global
evaluation item worded as “rate this instructor compared to others”). Thus,
researchers were essentially asking students to rank the qualities of good teaching and
then compared that with ratings results. For example, if students ranked stimulation
of interest, knowledge, friendliness, etc., as a very important quality, then an
instructor who rated high on those specific items should also rate high on “rate this
instructor compared to others.”

Dozens of studies appeared in the literature that often produced complex and
conflicting results. However, one of the most widely referenced authors on this topic
is Ken Feldman. In a series of articles published in "Research in Higher Education," Feldman (1976a, 1976b, 1977, 1978, 1979), synthesized the often contentious and sometimes confusing literature on student course evaluations. Feldman's work indicated that the qualities most closely associated with "superior college teaching" can be defined as stimulation of interest, clarity, preparation and organization, and impact of instruction. These characteristics, if presented through low-inference questions, provide dependable ratings results. Feldman's work has invariably been used as support for stakeholders who believe in the validity and reliability of student course evaluations. Parallel support for good teaching characteristics is seen in Marsh (1977), who found consistent ratings results between underclass students and graduating seniors, and Braskamp, Ory, and Pieper (1981), who found consistency between student ratings and student written comments.

Typical Faculty Concerns About Student Course Evaluations

Despite the evidence suggesting that student course evaluations are a valid measure of teaching effectiveness, and that they can reliably evaluate specific teaching traits, faculty are often skeptical. That skepticism, while mostly anecdotal, has largely been bolstered by some very famous (yet dubious) studies. These studies, in the face of much greater evidence to the contrary, have become the focal points of opposition to student course evaluations. Even though prior and subsequent research has essentially discredited them, these studies are noteworthy because of their lasting popularity. They continue to be referenced by higher education stakeholders to bolster opinion pieces regarding the perceived lack of validity and reliability of student course evaluations.
One of the original studies that gained early notoriety was by Rodin and Rodin (1972). In this study the authors reported a negative correlation (-.75) between student course evaluations and student achievement in an undergraduate calculus course. This study was reported in the journal “Science.” The study was accepted by critics of student course evaluations as prima facie evidence that student course evaluations do not correlate with student learning. The study is still referenced today. However, researchers have denounced the study because the evaluations were of teaching assistants, the measure of learning was not an end of semester exam but rather a cumulative total of correct answers throughout the term (each assistant also used different standards), and that this study stands out as an anomaly in the “ratings-achievement” literature. Marsh, Fleiner and Thomas (1975), who found substantial positive correlations between student ratings and performance, is one of many examples. However, the Rodin and Rodin (1972) study, subtitled, “students rate most highly instructors from whom they have learned the least,” (p. 1164), remains a hallmark study in the literature.

As famous as the Rodin and Rodin (1972) study is, the most well known study in the literature is known as “The Doctor Fox Effect.” Written by Naftulin, Ware and Donnelly (1973), this study has become the trademark for those who would like to cast doubt on the validity and reliability of student course evaluations. Known as “educational seduction,” Dr. Fox was an actor hired to lecture medical educators. Dr. Fox offered incorrect information, used fictional references, and otherwise lectured on content devoid of meaning for the target audience. Nevertheless, Dr. Fox was rated very highly, and this study gave rise to the still prevailing misconception that
high student ratings are the result of style rather than substance. The study has been widely criticized for its poor design, by Frey (1979), and Marsh and Ware (1982), among many others. The entire study was based on a single lecture, a one-time visit from a "doctor." His audience had no idea what he was talking about and had he remained longer than one lecture his ratings would certainly have changed. Ratings, in order to be valid, need to be administered at a time when students can make informed decisions, such as the end of the semester. Despite the problematic nature of everything about this study, critics of student course evaluations continue to cite it.

Using these studies as a basis to criticize the use of student course evaluations, publications continued to suggest that ratings are not valid or reliable. Writing in in the "Journal of Higher Education," Daniel Sheehan (1975), used the "biasing" argument, in which critics claim that there are so many ways that ratings results can be influenced, they should not be used in personnel decisions. However, Theall and Franklin (2001), Kulik (2001), and Ory and Ryan (2001), among many others provide logical arguments that student course evaluations can improve instruction and provide fair assessments for summative purposes. It is unreasonable, based on the weight of research evidence, to assert that ratings should not be a part of the faculty evaluation process. The very fact that virtually every college and university uses student course evaluations formatively and summatively supports their utility.

Dowell and Neal (1982) and Divoky and Rothermel (1988) concluded that class size, class type (required verses elective), and other variables could influence the outcome of ratings. There is some truth to the effect of course type. As Feldman (1978) indicates, "ratings tend to be somewhat higher for upper division courses and
elective courses, (and) ...humanities, fine arts, and languages tend to receive somewhat higher ratings" (p. 199). It is essentially common sense that an upper level, elective philosophy course would rate higher than an introductory, required chemistry course. A ratings system that cannot make an adjustment for that reality is doomed from the start.

Ambady and Rosenthal (1992) showed observers a 30-second-video clip (without audio) of professors. They called this "thin slices of expressive behavior," and found a .76 correlation between observer ratings and student course evaluation ratings of the same professor. Critics of student ratings have used this study to claim that student course evaluations are superficial and really do not measure anything meaningful. Of note in this study (similar to others like it), is that it only involved thirteen teachers and really represents an anomaly more than a standard. Feldman's (1989a,1989b) review of the literature, which synthesized over forty studies, found a positive correlation of .50 between student ratings and teacher self evaluations, peer reviews, former students, administrators and neutral observers. These raters not only heard what the teachers were saying but also were in the classroom during instruction. It is more convincing that these "thick slices" of teaching would yield more reliable results than the "thin slices."

Greenwald and Gilmore (1997a) tested a grading leniency bias. They analyzed the agreement between measures of student effort, grades and ratings results. They found a positive correlation between ratings and grades. Critics took their results and suggested that teachers who give out good grades get high ratings, and that teachers who give out poor grades get bad ratings. This study has often been
misinterpreted to suggest that teachers who are very demanding will receive poor ratings. Others have agreed with that sentiment, such as Gose (1997) and Perillo (2000), claiming that student course evaluations have caused grading leniency in higher education. Several ratings experts have questioned this study and others like it. March and Roche (1997), and McKeachie, (1997) point out that the major flaw with the "leniency" problem is that these studies do not remove the influence of course type. That is one of the biggest issues ratings experts have with the Greenwald and Gillmore’s piece. Almost any study that compares upper level elective courses to lower level required courses will yield a result that appears on the surface to suggest a grading leniency bias in student course evaluations, which is essentially the flaw in Greenwald and Gilmore's study.

Another very popular and widely cited article is from Williams and Ceci (1997), in “Change Magazine.” In "How'm I Doing: Problems with Student Evaluations and Courses," Williams and Ceci found that changes in speaking tone and glibness produced large changes in student ratings but not in test scores. This study is perhaps the essence of the debate on student course evaluations. The study focused only on one course and one teacher. It lacked all the multiple constructs of reliability and validity that sound research provides. However, the study has been widely cited. Perhaps because it appears in “Change,” or because it was from Cornell; whatever the reason, narrow research studies and opinion pieces continue to appear in journals and continue to raise doubts about the validity and reliability of student course evaluations. Liability concerns (Robinson and Fink, 1996) and charges of threats to academic freedom (Haskell, 1997a, 1997b, 1997c) also continue.
Establishing the Reliability and Validity of Student Course Evaluations:

Deconstructing

**Multi-Section Validity Studies**

While the current of opinion pieces against student course evaluations continue in higher education journals, the strongest evidence supports their reliability and validity. Researchers have taken four major approaches to establishing the reliability and validity of student course evaluations. The first major approach is known as the multi-section validity studies. This type of research evaluates the relationship between student ratings and student achievement in multiple sections of the same course taught by different professors. Researchers in these studies test the correlation between mean student ratings and mean student achievement on a common examination. Multi-section validity studies deconstruct two myths about student course evaluations; that they are not valid (myth #1) and they are not reliable (myth #2).

Marsh, Fleiner and Thomas (1975) found considerable correlations between mean student ratings and mean student achievement. This was one of the earliest studies that provided support for the validity and reliability of student ratings. An earlier study by Costin, Greenough and Menges (1971) which concluded that age, sex and level of student had little or no effect on ratings results, added further evidence that student ratings are valid and reliable regardless of the type of student that is completing the form. Cohen (1981) summarized forty-one multi-section studies and reported correlations of .40 to .55 between student achievement and various aspects of teaching (as measured by student course evaluations). Marsh (1987), in one of
several research studies he has published on student course evaluations, reports that course sections with the highest ratings also attain the best scores on standardized final examinations. Dozens of other research studies that involved hundreds of courses and thousands of students have also reported substantial correlations between ratings and achievement, for example, the Abrami, d’Apollonia, and Cohen, (1990) study. A fine and very current review of the validity and reliability of student course evaluations supported by multi-section studies can be found in Marsh and Roche (2000). Support for their overall efficacy in the teacher evaluation process, as developed through multi-section research, is found in Theall and Franklin (2001).

Multi-Method Studies

Multi-method studies test the convergent validity of student course evaluations by comparing them with other measures of teaching effectiveness. These studies test the correlation between ratings results and peer reviews, alumni, faculty self-evaluations, and others. Multi-method research attempts to attribute mean differences in student course evaluation outcomes to the instructor rather than characteristics such as student level, course type, or setting. Research has focused on a variety of courses and "results have generally shown evidence for the convergent and discriminant validity." (Ory and Ryan, 2001, p. 34). Multi-method studies help to deconstruct myth #3, that students are not qualified to make judgements about teaching competence and myth #4, that students are not able to make accurate assessments of teaching until after they have graduated.

Feldman (1989a, 1989b) provides the best syntheses of multi-method studies. He found between .15 and .42 correlation between student ratings and other
observers. Marsh (1987) found a .49 agreement between student ratings and teacher self-evaluations. Marsh (1982) also supports multi-method research for validity. However, the strongest support for multi-method research is found in Murray (1983). In this study, strong evidence, from a variety of observers, concluded that student course evaluations could reliably measure nine specific teaching behaviors. Those nine items, such as clarity and organization, are consistent with those found by Feldman (1976a, 1976b). The vital component of Murray's work focused on making the questions "low-inference," a type of question in which the student only has to decide how a specific teaching behavior affected her or him individually. Thus, low-inference questions have very high reliability, are consistent across a variety of teaching measures and are positively related to observable teaching behaviors.

**Bias Studies**

Bias studies test the influence of certain variables on the outcomes of student course evaluations. These extraneous variables include teacher personality, grades, workload, leniency and course type. The research exposes few biasing influences on student course evaluation outcomes. While it has been established that there is some influence on course type (introductory verses elective), it is much more the case that the instructor's effectiveness (or lack thereof) is the reason that ratings are high or low.

Students' expected grade was tested by a variety of researchers but some of the larger studies include Centra and Creech (1976), Marsh (1987), and Feldman (1997), who found low correlations (.10 to .20). Accordingly, Abrami et al., (1980) found no statistical significance between grading standards and ratings outcomes.
Likewise, researchers found a small relationship between earned grades and ratings results (.10 to .30). Higher grades do not naturally result in higher ratings (Theall, Franklin and Ludlow 1990, Feldman 1997, Marsh and Roche 2000)

Workload is often cited as a potential bias in student course evaluations. The theory is that easier teachers will receive higher ratings and harder teachers will receive lower ratings. The opposite relationship has been found. Marsh (1987), Cashin (1988), Centra (1993), and Franklin and Theall (1996) all found that workload was positively correlated with ratings. Finally, bias theory invariably claims that a teacher's personality (regardless of content knowledge or student learning) will result in high ratings. In reality, an effective teacher is one who engages his or her students, is enthusiastic and models a love of learning. Those attributes motivate students to learn; it is not "personality" that gets high ratings, but good teaching. Popular teachers may in fact be popular because they are good teachers. As Theall and Franklin (2001) state, "The assumption that popularity somehow means a lack of substance or knowledge or challenge is totally without merit. The are no studies to support this view." (p. 53). Taken collectively, bias studies deconstruct myth # 5, that ratings are popularity contests, myth # 6, students only give high ratings if they receive high grades, and myth # 7, extraneous variables aversely affect ratings outcomes.

**Meta-Analytic Studies**

Meta-Analytic studies synthesize the major findings of research on student course evaluations. Results from this type of research have been reported in narrative as well as analytic form. Ultimately, these studies attempt to describe a shared set of
dimensions that can be accurately measured by student course evaluations. Together, meta-analytic reports serve as summary evidence for the reliability and validity of student course evaluations.

Feldman (1977, 1978) provide important considerations for the use of student course evaluations. Within these two publications, Feldman analyzed dozens of studies and his results suggest that certain "operating procedures" central to the administration of ratings forms will help ensure their reliability and validity. Feldman's research suggests that the conditions under which student course evaluations are administered needs to be standardized across the entire university or college. Those conditions include: whether the students remain anonymous or identify themselves, if the students are aware of the intended use of ratings results and if the instructor is present at the time of administration. Feldman states that however these conditions are handled, they must be handled the same way each time, and consistently across campus. For example, when students know results are for promotion, or when the instructor remains in the room, the results are higher ratings. Thus, anonymity, intention and instructor presence in the room at the time of administration should be handled the same way for each course. Additionally, Feldman found no impact on the timing of administration.

Cohen (1981) is a research study that is often cited to support the validity and reliability of student course evaluations. Cohen synthesized forty-one prior studies, using final examination scores to measure achievement. The average correlation between achievement and ratings was .43, and the average correlation between achievement and overall rating of the course was .47. Even higher correlations
(above .50), were found between achievement and certain low-inference teacher characteristics such a clarity and organization. Correlations were almost zero for items that centered on course difficulty. Certain other generalizable components of student course evaluations were found. Those include higher correlations in courses taught by full-time faculty and examinations graded externally. All of these results support the validity and reliability of student course evaluations.

Yet, faculty concerns remain evident in the literature. Three of those concerns, teaching controversial topics, student work habits and faculty gender have been addressed by meta-analytic research. Marsh (1987) did not reveal any difference in ratings results, as they relate to "controversial topics." Faculty have been known to complain about ratings because they believe that students who do not perform well will blame the teacher, which will result in low ratings. Theall, Franklin and Ludlow (1990) studied student "attributions of performance" to test if students attributed their academic performance to the teacher. They found that students attributed their own performance to their expected and earned grade, and that this had no significant impact on ratings results. Franklin and Theall (1994) found no significant difference in ratings results by gender of faculty. However, if a department systematically assigns females to only introductory level required courses (as opposed to higher level elective courses), ratings results will be invalid and that example is a potential contaminant of any ratings system.

Student Course Evaluations: Designing a Best Practice Approach

As the weight of research evidence suggests, student course evaluations are valid and reliable. They are also very useful for improving instruction and as part of
personnel decisions. The literature also reveals two often-opposing realities about student course evaluations. The first is that the use of student course evaluations is virtually ubiquitous in higher education. The second is that opinion pieces (and flawed research) continue to gain headlines in higher education publications. It seems odd that faculty and even some administrators do not endorse student course evaluations while their psychometric properties are valid, reliable, and stable across a variety of teaching dimensions. It seems plausible that the apparent reason for this contradiction is found within the various "local" applications of ratings systems. For a ratings system to be effective, its users must be educated in three primary knowledge components of ratings use.

Knowledge of Obtaining Data

To design and implement an effective ratings system the first knowledge component users need to be familiar with is obtaining data. Obtaining data refers to who, what, when and how of collecting data from student course evaluations, based on research evidence found in the literature. Effective teaching is multi-dimensional, and it is therefore important for users to keep focused on highly correlated, stable dimensions of effective teaching in designing a ratings system.

As Cashin (1990) reports, faculty in science, engineering and math usually receive lower overall ratings than faculty in fine arts, humanities and health-related professions. Recognizing this, designers of ratings systems should heed Franklin and Theall (1995), who state: "...it is inadvisable to compare instructors or courses across disciplines." (p. 44). In this same study Franklin and Theall did find a relationship between how "valuable" students viewed their assignments from professors and
ratings results. Correlations up to .65 in some disciplines were found, suggesting that the more valuable students feel their time spent preparing for class, the higher they will rate their professors. Ratings systems designers should include a "time-valued variable" as a question on student course evaluations. The time-valued variable is important because it is reliable across all disciplines.

As previously mentioned, conclusions have been strong about stable dimensions of teaching (Feldman 1976a, 1976b; Cohen, 1981; Marsh, 1987). Users should also be familiar with Abrami and d'Apollonia (1990), who identified twenty-four teaching dimensions that are stable, and low-inference. Cashin and Downey (1992) suggest two or three items that ask students to rate general satisfaction with the course and the amount they have learned.

Together, these studies and the evidence about student course evaluations they produced should guide local formation of questionnaires. A best practice approach would include student course evaluation questionnaires that ask low-inference questions (Murray, 1983). The questions should be based on stable teaching dimensions. Data should be collected from courses within the same discipline. There should be "global" items that ask students to "rate this instructor compared to others," and general satisfaction with the course. It appears however, that ratings designers are not as familiar with the literature as needed. Ory and Wieties (1991), in a study of evaluation items chosen by faculty, found faculty choose questions that were of little value and not connected to the literature. Franklin and Theall (1989) found a lack of knowledge with the literature that correlated with unfavorable attitudes toward student course evaluations. Perhaps even worse, Franklin and Theall (1991) report
that faculty will deny or ignore research results. Evidence supports the potential for effective formative and summative student course evaluation designs. However, it is apparent that users are either unfamiliar with or unwilling to listen to that evidence. Thus, many users are not knowledgeable enough of the literature on obtaining data.

Knowledge of Interpreting Data

Knowledge of interpreting data has to do with the ability of users to apply, run and analyze the statistical results of student course evaluations. There are several important considerations within "interpreting results." To begin, users must be knowledgeable with measures of central tendency such as mean, median and mode. Centra (1979) points out that evaluation results will normally include (and users should be able to interpret) analyses such as standard deviations, t-tests, anovas, and percentile ranks. Additionally, an awareness to run and interpret item analyses, especially global items, is vital to successful interpretation.

Ratings experts point out that knowledge of interpreting data is especially lacking. Thompson (1988) points out that unless the data is very simple, most users will fatally misinterpret the results. Doyle (1989) notes that academics that work with ratings usually have a weak statistical background and suffer from "paralysis of analysis." Franklin and Theall (1989) found significant lack of interpretation knowledge among faculty. Theall and Franklin (1991) note that even users who are competent with statistics will sometimes misinterpret results. For instance, student comments generally agree with quantitative results, but to validate their comments, seventy-five percent of respondents should be writing the same thought. That is also true of the percentage of respondents filling out forms. Seventy-five percent is
needed in a class to validate results. Knowledge of percentile ranking is also faulty in many institutions. Franklin and Theall (1999), in a similar study to their (1989) research, found that "...faculty vary widely in their ability to understand and interpret statistical summaries of their students' responses" (p. 221). Knowledge of interpreting data seems to be lacking throughout all levels of current ratings systems.

Knowledge of Communicating Data

Knowledge of communicating data refers to making decisions based on the results of data produced through student course evaluations. These decisions take many forms. They include face-to-face consultation, printed reports, written reviews and any other way of disseminating the results of ratings. Communicators of results must be cognizant of striking the right balance. As Miller (1987) points out, ratings should be used as part of a comprehensive assessment process. Ultimately, when personnel decisions include tenure, promotion, retention, and merit pay, individuals who make those judgements are truly the evaluators of teaching effectiveness. Such determinations bear great responsibility, and carry the implication that those who make them need to be familiar with the literature and skillful in the three major knowledge components of student course evaluations.

Unfortunately, the literature reveals that too many decision-makers are not as familiar or skillful as they need to be. Theall and Franklin (1990a) state, "many who make personnel decisions, based at least partly on student ratings, are unfamiliar with the most accepted literature in the field" (p. 24). Franklin and Theall (1999) report that the communicating of results, while the most important component of any ratings process, is often taken for granted and is the least understood. Cranton and Smith
(1986) demonstrated the challenge of extending norms across an entire campus. For example, the literature is very clear about avoiding comparisons between course types. Yet, Theall and Franklin (2001) report of administrators asking not only for university norms but "to produce reports of average scores to the third decimal point" (p. 2).

Knowledge of communicating data is critical to the success of any ratings system. As Erickson and Erickson (1979) point out, improving teaching and making personnel decisions is not one and the same. Student course evaluations can be effective parts of both formative and summative processes. The literature suggests that users are not knowledgeable enough in the complex tasks that are associated with designing, implementing and making informed judgements regarding ratings systems. As Franklin and Theall (1990) suggest, "Fundamental safeguards and standards for practice are easy to overlook when the system designer assumes that questionnaires and ratings systems are synonymous" (p. 82). Much of the evidence in the literature suggests that users are not knowledgeable enough in the area of communicating data.

Summary of Literature Review

Student course evaluations have grown steadily in their use and utility. Ratings results from student course evaluations have tremendous benefit for teaching improvement and as part of personnel decisions. The weight of research evidence suggests that student course evaluations are valid, reliable and can accurately measure a variety of teaching dimensions. Four major research methods, multi-section, multi-method, bias and meta-analytic studies have provided the most generally recognized frameworks for their validity and reliability. Still, opinion columns persist in higher
education journals. Moreover, a few dubious studies are continually cited, and they help to perpetuate seven common myths about student course evaluations.

In deconstructing those myths, the literature bears upon three knowledge and skill areas essential to establishing a best-practiced approach to using student course evaluations as part of a comprehensive ratings system. Users should be educated in the knowledge of obtaining data. This includes what types of questions to ask on student questionnaires, how to administer evaluations, and establishing valid and reliable data collection, processing and analysis tasks. Users also need to be knowledgeable in interpreting data. This knowledge is the ability to decipher the quantitative outcomes of ratings. Finally, users need to be knowledgeable in communicating data. This includes the ability to place ratings results in the appropriate context, understanding the relative weight of each evaluation criteria assigned by the department or university and making fair personnel decisions. Research indicates that many users lack the requisite knowledge in one of more of these three areas.
CHAPTER III

METHODOLOGY

Research Design

This study employed a quasi-experimental design to test the effect of educational interventions on faculty and administrator knowledge and attitude toward student course evaluation variables. Two distinct interventions were tested in the research. The research was designed to compare the knowledge and attitude change of respondents depending on the type of intervention. The two experimental groups and a control group were administered a pre and post survey (same survey).

The Workshop Group was given the survey to complete upon their arrival at a March 30, 2001 Colloquium on student course evaluations, held at Seton Hall University (Colloquium, 2001). At the conclusion of the workshop, participants were again asked to complete the survey. The purpose of this was to test if any change took place over time in respondents’ knowledge or attitude regarding student course evaluations as a result of attending a workshop on the key student course evaluation knowledge variables and best practice.

The Reading Group was asked to complete the pre-survey and mail it in a pre-paid envelope. After placing the pre-survey in the mail, participants were asked to read three articles: Franklin & Theali (1990b), Theall & Franklin (1990a) and Theali & Franklin (1991). After completing the readings participants were asked to mail the
post-survey in a pre-paid envelope. The purpose of this was to test if any change took place in respondents' knowledge or attitude regarding student course evaluations after reading three articles on the key student course evaluation knowledge variables.

The third group constituted a "no-intervention" or Control Group. These participants were asked to complete the pre-survey. When the survey was returned, they were asked to complete the post-survey. The Control Group did not receive any type of educational intervention between the two surveys. The purpose of the Control Group was to test if any change took place over time in respondents' knowledge or attitude regarding student course evaluations, as a result of subject maturation or other extraneous features.

Study Participants

Population

The population targeted for this study were higher education faculty and administrators in New Jersey. The literature suggests that those two stakeholders are most closely linked with student course evaluation outcomes. Faculty teaching effectiveness is measured by student course evaluations, and depending on the institution, could be the most important consideration in promotion and tenure. Administrators often help to create forms, interpret results, and make personnel decisions based on student course evaluation outcomes. Thus, faculty and administrators are the focal participants of this study.

Sample

The sample in this study consisted of ninety-three participants who completed the pre and post survey. Demographic information was requested in questions 1-5 of
the survey that included rank, experience, duties, participation in personnel decisions and participation in creating or selection forms. Figures 1 through 5 illustrate the demographic breakdown of study participants.
Figure 1. Faculty Rank

Figure 2. Years of Experience in Higher Education
Figure 3. Current Duties

Figure 4. Level of Participation in Personnel Decisions
**Figure 5.** Level of Participation in Creating or Selecting Forms
Thirty-three participants were in the Workshop Group and thirty each were in the Reading and Control Groups. Sixty-eight were faculty and twenty-five were administrators, who came from twenty-six colleges and universities in New Jersey. Seven state, ten private colleges and universities and nine community colleges are represented in the study, as presented in Figures 6-7. The three groups were all recruited and selected by contacting the chief academic officer at each school.
Figure 6. Participation by Group

Figure 7. Colleges and Universities Represented
The Workshop Group was recruited to attend a colloquium on "current issues in higher education." Every chief academic officer (CAO) in New Jersey public, independent and community college received a letter about the colloquium (Appendix A). In the letter CAOs were asked to identify up to four faculty or administrators who would benefit from a workshop on student course evaluations. The letter was then attached as a file and sent by email to the CAOs who did not respond to the hard copy. A third contact was made by phone to all those CAOs who did not respond to the letter or email. Individuals who expressed interest were sent registration materials (Appendix B), to be returned to Dr. Martin Finkelstein, Seton Hall University, who supervised the study. In all thirty-nine people attended the March 30, 2001 colloquium. Of those thirty-nine, three comprised a "panel of experts," three were full time faculty at Seton Hall, and the other thirty-three comprised the Workshop Group participants.

The Reading Group was recruited by first writing a letter to every CAO in New Jersey inviting them to identify up to four people they thought would benefit from participating in a study on student course evaluations (Appendix C). The letter was then sent as file attachment to every CAO who did not respond to the hard copy. A third contact was made by phone to those CAOs who did not respond to the letter or email. Individuals who expressed interest were sent a packet of materials that included a cover letter with instructions (Appendix D), two pre-paid envelopes, two surveys (pre and post), and the three readings (Appendix E, F, G). In all, thirty participants returned pre and post surveys.
The Control Group was recruited by asking the CAOs to please identify any faculty or administrators they thought would be interested in participating in a continuation of a study on student course evaluations. The letter was sent as a file attachment (Appendix H). Individuals who expressed interest were emailed the pre survey. When the survey was returned electronically, respondents were then sent the post survey. In all, thirty participants completed the pre and post survey.

Educational Interventions

There were two active educational interventions in this study. The first intervention was given to the Workshop Group. The intervention was held at Seton Hall University on March 30, 2001. The intervention came in the form of a workshop that began at noon and finished at 3 PM. The presenter for the day was Dr. Mike Theall, University of Illinois. Dr. Theall is a noted expert on the psychometric properties of student course evaluations. He is widely published on the topic and provided participants with a best practice approach to designing a student course evaluation ratings system. Dr. Theall used overheads, Power Point slides, lecture, and guided practice during his three hours. Participants were offered the opportunity to complete the survey upon their arrival and at the conclusion of the workshop.

The second intervention was given to the Reading Group. This intervention came in the form of three articles published by workshop presenter Mike Theall (and co-author Jennifer Franklin). They were Theall and Franklin (1990a) (see Appendix E), Franklin and Theall (1990) (see Appendix F) and Theall and Franklin (1991) (see Appendix G). Participants received the pre survey by mail and were instructed (see Appendix D) to complete it and mail it back to the researcher in a pre-paid envelope.
After mailing the pre survey participants were asked to open the envelope that contained the three articles, and read them. Participants were then asked to complete the post survey and mail it back to the researcher in a pre-paid envelope.

The Control Group did not receive any intervention. Participants received the survey in electronic form as an email file attachment. They were asked to complete the survey and returned it to the researcher electronically (they did so by clicking "submit"). When the researcher received the pre survey, participants were sent the post survey. They were asked to complete the post survey and returned it to the researcher electronically.

Hypotheses

The study sought to test the following hypotheses:

1. At T1, subjects with higher knowledge will demonstrate significantly more favorable attitudes towards student course evaluations.

2. At T2, subjects with higher knowledge will demonstrate significantly more favorable attitudes towards student course evaluations.

3. At both T1 and T2, administrators will demonstrate significantly more knowledge and significantly more favorable attitudes toward student course evaluations.

4. Between T1 and T2, those subjects afforded more intensive educational experiences will demonstrate significantly higher knowledge and favorability (Workshop Group) than those subjects offered less intensive (Reading Group) or no education at all (Control Group). For the (Workshop Group), knowledge will increase significantly more than the (Reading Group) and will result in a significant increase in attitude.
5. For the Reading Group, knowledge will increase more moderately from T1 to T2 than the (Workshop Group) but greater than the (Control Group).

6. For the (Control Group), knowledge will not significantly increase from T1 to T2 and there will not be a significant change in attitude.

Data Collection

Conceptualizing the Variables

Obtaining Data: knowledge variables as they relate to obtaining data will focus on familiarity with the results of research studies conducted on the reliability and validity of student course evaluations. Examples of this include level of agreement between student course evaluation outcomes and other measures of teaching and learning, the proportion of a class needed to validate results, and the relationship between grades and ratings outcomes. Since the weight of empirical evidence suggests that the results of these investigations are clear, users should be cognizant of them.

Interpreting Data: knowledge variables as they relate to interpreting data are the ability to use statistics and statistical information. For example, users must be familiar with standard deviation, understanding that a 1.2 or less generally indicates agreement. Other related statistical knowledge includes mean scores, percentile ranking, and t and z scores. These variables will comprise the interpretation component of this study.

Communicating Data: variables as they relate to communicating data focus on the ability the use knowledge of obtaining and interpreting data for best practice. Deans, chairs, and faculty alike must be able to make informed decisions regarding
ratings data. Users must be able to advise and make personnel recommendations based on student course evaluation outcomes. As an example, users need to be informed about the difference between statistical feedback and written comments. Another example is the knowledge of how many course evaluations are needed to make them valid.

Attitude: variables that include a favorable or unfavorable position on items related to student course evaluations. Two of these include a student’s right to rate courses (and instructors), and whether or not students can provide accurate responses to questions about teaching and learning. Another attitude variable relates to whether or not a course or an instructor’s teaching can be improved based on information in student ratings.

Operationalizing the Variables

The conceptual framework for his study is built around four knowledge variables of the teaching evaluation process found in the literature. Those variables are obtaining data, interpreting data, communicating data, and attitude toward student course evaluations. To operationalize these variables a survey on student ratings was administered. Initially, respondents completed demographic information related to rank, years of teaching, level of participation in personnel decision making, current duties, and whether or not they have ever helped create ratings forms. The first five questions of the survey were demographic questions.

The survey contains questions that targeted each of the four variable sets of the study. There were fifteen questions that asked respondents to react to statements about knowledge and five attitude questions. The statements reflect principals that
should guide any best practice in the use of student course evaluations. All knowledge questions were accompanied by six choices for each statement: strongly agree, tend to agree, tend to disagree, strongly disagree, or uncertain. Therefore, respondents who are familiar with research evidence related to student course evaluations should correctly affirm or negate various knowledge statements. The survey also asks questions in which respondents need to be familiar with quantitative subject matter regarding student course evaluation statistical data.

Knowledge of obtaining data is linked to essential variables regarding student course evaluation best practice. Users must be cognizant of data produced through research. For instance, users should be aware that global, low inference items are the most valid and reliable measures found from student course evaluation research. A user familiar with the literature should disagree with this statement: "very specific items dealing with a teacher's behavior and skills are usually more reliable than broad, global items such as this instructor compared to others." As another example, respondents should disagree with the idea that the total number of student raters is more important than the percentage of a class that rates a professor (both are essential).

Knowledge of interpreting data is based on a respondent's ability to correctly interpret statistical outcomes related to student course evaluations. In this section of the survey, participants were asked questions related to standard deviation, percentile rank, ratings scales, range and direction, along with t and z scores. Respondents were asked to react to statements regarding teaching effectiveness, based on their knowledge of quantitative student course evaluation results. Familiarity with the
literature, in addition to skill in quantitative analysis is essential to the interpreting variable.

Knowledge of communicating data is also linked to student course evaluation best practice. Qualitative concerns surrounding what users need to know when making decisions based on ratings results is the cornerstone of this variable. For instance, respondents should agree that available, relevant comparison groups should be considered in personnel decisions. Another example is that respondents should agree that the percentage of respondents should be factored for each choice on a particular item. Ultimately, the user must be aware of several communication variables regarding student course evaluations to ensure best practice. The survey tests the extent of respondent knowledge regarding communication.

Attitude is tested through reaction to statements regarding the respondents' experience with student course evaluation use. The following statements, in which participants are asked to agree or disagree (six choices), are examples of how attitude as a variable is tested:

"Specific questions about teaching behaviors in student course evaluations can help me improve my teaching."

"Students generally provide sincere, honest and accurate responses to questions about teaching effectiveness."

"Student ratings questionnaires are not a useful way to measure teaching effectiveness."
Instrument

The original survey (Appendix I) was constructed as part of research conducted by Franklin and Theall (1989). The original survey was established through published research findings on student course evaluations. That research focused on the reliability and validity of student course evaluations as an accurate measure of teaching effectiveness and student learning. Thus, the survey questions were crafted around the specific language found in those publications, the vast majority of which have been mentioned in chapter one and chapter two. Franklin and Theall took the results of those studies and combined them with the results of their own research to produce the original survey (Appendix I).

Once they had created their survey they sought to validate it by having it reviewed by the very experts on whose research it was based (Appendix J). Some changes were made and that panel of experts agreed upon the "correct" responses. Thus, the survey questions were based on published results of qualified research, and were validated by a panel of experts that produced that research.

The survey used in this study is a shorter version of the original with some slight modifications. Several of the questions are verbatim from the original. Franklin and Theall (1989) used the original survey to conduct research that culminated in a paper presented at the annual meeting of the American Education Research Association (AERA). The focus of that presentation was the relationship between knowledge and attitude. They found significant associations between knowledge and attitude and also found a wide discrepancy between the knowledge of experts and that of the 779 respondents to the survey. In addition to establishment of
the construct validity of the original survey (through the expert group), Franklin and Theall (1989) utilized an alpha scale test the reliability of each of the knowledge domains and attitude questions.

On the survey used for this study (Appendix K), questions 11, 17, 23, and 24 have been modified. Further research and publications were studied to create the best possible survey and also to produce an instrument that could measure the particular variables of this study while taking less time to complete (see Appendices E, F, and G). Three articles by the authors provided the framework for the modifications. For instance, questions 11 and 17 are taken from Theall and Franklin (1991). Modifications to questions 23, and 24 are taken from Franklin and Theall (1990).

Questions in this survey were worded to match the publications mentioned in this section. Those articles closely relate to the variables being studied and measured in this study. Each question has been validated by research studies conducted by the experts who examined the construct of the original survey. They also match the variables that comprise the focus of this study. Questions 1-5 ask the respondent for demographic information. Questions 6-10 ask knowledge questions based on obtaining data. Questions 11-15 ask knowledge questions based on communicating data. Questions 16-20 ask knowledge questions based on interpreting data. Questions 21-25 seek respondents' attitude toward student course evaluations.

Data Collection Procedures

Data were collected in three stages aligned with the three groups in the study. Collection for the Workshop Group was facilitated through attendance at a Professional Colloquium on student course evaluations held at Seton Hall University
on March 30, 2001, from noon to 3 PM. In January, letters of invitation to the
Colloquium were sent to every Chief Academic Officer at all community, private and
state colleges and universities in New Jersey (Appendix A). Through contact with the
Chief Academic Officer(s), eighteen colleges were represented at the Colloquium.
The Colloquium participants completed the pre survey when they arrived and the post
survey at the conclusion of the presentation.

Collection for the Reading Group was facilitated through mailed survey
response. Chief Academic Officers were again contacted to determine if faculty or
administrators would be willing to participate in a continuation of the study that
began with the Colloquium (Appendix C). As before, the Chief Academic Officer(s)
disseminated the information regarding the study and individuals who were interested
in participating contacted the researcher. Those individuals were sent a "packet of
information" to guide their participation. Participants were instructed to complete the
pre survey on the outside of the package and mail it to the researcher (Appendix D).
They were then instructed to open the package, read the three articles enclosed,
complete the post survey, and mail it to the researcher.

Collection for the Control Group was facilitated by email correspondence.
Initially, an email invitation to participate in a continuation of the study was sent to
the same Chief Academic Officers (Appendix H). After they had disseminated the
information, individuals that wanted to participate sent the researcher an email
indicating their interest. When received, each individual was emailed the pre survey.
Participants who completed the pre survey sent it to the researcher via email, and
were then sent the post survey (also via email). Respondents then sent the post survey to the researcher.

Data Analysis

To test the hypotheses, data was analyzed using several statistical techniques. Frequency distribution tables were generated to present the demographic data of participants (as shown in figures 3.1-3.5). Three group variables consisted of the Workshop Group, the Reading Group and the Control Group. Using the syntax commands within SPSS, knowledge and attitude variables were created for pre and post survey responses. Variables one and three comprised pre and post knowledge responses. Variables two and four comprised pre and post attitude responses. Variable five consisted of the participants who identified themselves as administrators and variable six consisted of the participants who identified themselves as faculty.

To determine level of knowledge, participant responses were given a numerical value. A five-response lickert scale was used to guide data entry. If a respondent's choice on a question represented the "highest knowledge" response, data entry for that question became a five. That would indicate that the respondent was most familiar with the literature, as it related to each particular question. Responses to knowledge questions were coded on a five point scale, from 5 = most knowledgeable, 4 = knowledgeable, 3 = Uncertain, 2 = less knowledgeable 1 = not knowledgeable. Accordingly, the five attitude questions were coded on a five point scale from 5 = most favorable, 4 = favorable, 3 = Uncertain, 2 = less favorable 1 = not favorable.
The variables "pre knowledge" and "post knowledge" were computed by summing participant knowledge scores on questions 6-20 of the pre and post survey. Pre knowledge was variable one post knowledge was variable three. On each of the knowledge questions, either "strongly agree" or "strongly disagree" represented the "most knowledgeable" response, depending on the question. Strongly agree represented the most knowledgeable response on questions 7, 13, 14, 16, 19, 20. On those questions the numeric equivalent of strongly agree = 5. Accordingly, tend to agree = 4, uncertain = 3, tend to disagree = 2 and strongly disagree = 1. Conversely, on questions 6, 8, 9, 10, 11, 12, 15, 17, 18, strongly disagree represented the most knowledgeable response. On those questions the numeric equivalent of strongly disagree = 5, tend to disagree = 4, uncertain = 3, tend to agree = 2, and strongly agree = 1.

Summing participant attitude scores on questions 21-25 of the pre and post survey created the variables “pre attitude” and “post attitude.” Pre attitude was variable two and post attitude was variable four. On each of the attitude questions, either "strongly agree" or "strongly disagree" represented the "most favorable" response, depending on the question. Strongly agree represented the most favorable response on questions 22, 23, 24, 25. On those questions the numeric equivalent of strongly agree = 5. Accordingly, tend to agree = 4, uncertain = 3, tend to disagree = 2 and strongly disagree = 1. Conversely, on question 21, strongly disagree represented the most favorable response. On that question the numeric equivalent of strongly disagree = 5, tend to disagree = 4, uncertain = 3, tend to agree = 2, and strongly agree = 1. The numeric equivalents of all knowledge and attitude questions are represented
by (Appendix L). The mean attitude and mean knowledge score of each group
(Workshop, Reading and Control) was calculated on the pre survey and then
calculated again on the post survey. Likewise, the mean attitude and mean
knowledge score of the administrator group and the faculty group was calculated on
the pre survey and then again on the post survey.

Hypothesis one and two of this study tested the premise that subjects with
higher knowledge will demonstrate more favorable attitudes toward student course
evaluations. The first analysis of this study centered on the relationship between
knowledge and attitude. Using variable one and variable two, a Pearson Correlation
was used to test the hypothesis that higher knowledge will be correlated with a more
favorable attitude in the pre survey. Using variable three and variable four, a Pearson
Correlation was used to test the hypothesis that higher knowledge will be correlated
with more favorable attitude in the post survey.

Hypothesis three of the study tested the premise that administrators will
demonstrate more knowledge and more favorable attitudes than faculty toward
student course evaluations. Using dependent variables one, two, three and four and
independent variables five and six, an ANOVA test was run to see if significant
differences in knowledge or attitude were found between administrators and faculty.
The ANOVA tested mean knowledge score differences on the pre survey and mean
knowledge score differences on the post survey. The ANOVA also tested mean
attitude score differences on the pre survey and mean attitude score differences on the
post survey.
Hypothesis four of the study tested the premise that subjects afforded a more intensive educational experience will demonstrate higher knowledge and a more favorable attitude toward student course evaluations. Mean scores were calculated for the Workshop Group, the Reading Group and the Control Group, and the descriptive results were presented. Within this test, the three groups comprised the independent variables while pre and post knowledge, along with pre and post attitude comprised the dependent variables. An ANOVA was used to test for significant differences between the groups in knowledge or attitude. The ANOVA tested for significant differences on both the pre and post survey. To further test which group knowledge or attitude scores had significant differences, a Least Significant Difference was utilized, identifying individual difference between groups.

Hypotheses five and six of this study tested if significant increases in knowledge and attitude resulted within each group, as a result of the educational interventions. Hypothesis five and hypothesis six tested the change, within each group, of the mean knowledge score on the pre survey from the mean knowledge score on the post survey. These two hypotheses also tested the change, within each group, of the mean attitude score on the pre survey from the mean attitude score on the post survey. The mean score on knowledge in the pre survey was subtracted from the mean knowledge score in the post survey. Likewise, the mean attitude score on the pre survey was subtracted from the mean attitude score on the post survey.

Three tests were used to analyze the mean score changes in knowledge and attitude. To initially test the significance of change on the mean knowledge and attitude scores, a Paired Samples test was used. This test analyzed the change in
knowledge and attitude within each group. Second, an ANOVA test was used to analyze the change between the groups. The ANOVA tested the overall significance of change between the Workshop Group, the Reading Group and the Control Group. Third, a Least Significance Difference was used to individually itemize the significance of change between the three groups.

Limitations

The biggest limitation of the study is the sample size. In most research on student course evaluations, such as the multi-section validity studies, the (N) was in the thousands and some are in the tens and hundreds of thousands. The cost of materials and postage was a constraint. This study, combining all three groups, will be less than 100. As a result, it will be hard to generalize to all of higher education. The study would most appropriately be replicated in a specific division, department, or college that was part of a larger university.

A second limitation of the study is that the knowledge and attitude being measured are short term. The effects of the educational intervention for the workshop group were tested after a period of three hours (before and after the Colloquium). The effects of the educational intervention for the reading group were tested after 1-3 weeks (before and after the readings were completed). The effects of the control group were likewise tested after 1-3 weeks. Thus, there was no mechanism to measure if the knowledge increase and/or change in attitude were retained over longer periods.

A third limitation is sample selection. Participants were selected through initial contact with the CAO of each college and university. Three attempts were
made (mail, email and phone). Some COAs did not respond and some colleges had no representation due to schedule constraints. As a result, the sample, due to the selection process for this study was not completely random.
CHAPTER IV
DATA ANALYSIS AND FINDINGS

Sample Characteristics

The purpose of this study was to test the effect of an educational intervention on faculty and administrator knowledge and attitude to student course evaluations. The study consisted of ninety-three participants, who were asked five demographic questions on the survey (Appendix K). The demographic breakdown of participants is presented in Tables 1-5.

Table 1

Sample Characteristics: Rank

<table>
<thead>
<tr>
<th>Rank</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Professor</td>
<td>20</td>
<td>21.5</td>
</tr>
<tr>
<td>Associate Professor</td>
<td>19</td>
<td>20.4</td>
</tr>
<tr>
<td>Assistant Professor</td>
<td>24</td>
<td>25.8</td>
</tr>
<tr>
<td>Instructor</td>
<td>12</td>
<td>12.9</td>
</tr>
<tr>
<td>Adjunct</td>
<td>8</td>
<td>8.6</td>
</tr>
<tr>
<td>Non-Faculty</td>
<td>10</td>
<td>10.8</td>
</tr>
</tbody>
</table>
### Table 2

**Sample Characteristics: Experience**

<table>
<thead>
<tr>
<th>Experience</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;3</td>
<td>4</td>
<td>4.3</td>
</tr>
<tr>
<td>3-6</td>
<td>13</td>
<td>14.0</td>
</tr>
<tr>
<td>&gt;6-9</td>
<td>14</td>
<td>15.1</td>
</tr>
<tr>
<td>&gt;9-12</td>
<td>11</td>
<td>11.8</td>
</tr>
<tr>
<td>&gt;12-15</td>
<td>6</td>
<td>6.5</td>
</tr>
<tr>
<td>&gt;15</td>
<td>45</td>
<td>48.4</td>
</tr>
</tbody>
</table>

### Table 3

**Sample Characteristics: Current Duties**

<table>
<thead>
<tr>
<th>Current Duties</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teaching Only</td>
<td>34</td>
<td>36.6</td>
</tr>
<tr>
<td>Teaching with Administrative</td>
<td>34</td>
<td>36.6</td>
</tr>
<tr>
<td>Administrative Only</td>
<td>11</td>
<td>11.8</td>
</tr>
<tr>
<td>Administrative with Teaching</td>
<td>14</td>
<td>15.1</td>
</tr>
</tbody>
</table>

### Table 4

**Sample Characteristics: Participation**

<table>
<thead>
<tr>
<th>Participation</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Often</td>
<td>47</td>
<td>50.5</td>
</tr>
<tr>
<td>Sometimes</td>
<td>22</td>
<td>23.7</td>
</tr>
<tr>
<td>Never</td>
<td>24</td>
<td>25.8</td>
</tr>
</tbody>
</table>
Table 5

Sample Characteristics: Creating or Selecting Forms

<table>
<thead>
<tr>
<th>Creating or Selecting</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Created Forms</td>
<td>36</td>
<td>38.7</td>
</tr>
<tr>
<td>Chose Forms</td>
<td>9</td>
<td>9.7</td>
</tr>
<tr>
<td>Never Created Forms</td>
<td>17</td>
<td>18.3</td>
</tr>
<tr>
<td>Never Chose Forms</td>
<td>31</td>
<td>33.3</td>
</tr>
</tbody>
</table>

Question one asked participants to circle their faculty rank, the results being indicated in (Table 1). Question two asked participants to circle the number of years they had work in higher education, the results being indicated in (Table 2). Question three asked participants to circle their current duty, the results being indicated in (Table 3). Question four asked participants to circle their level of participation in personnel decisions in which student ratings were offered as evidence of teaching performance, the results being indicated in (Table 4). Question five asked participants to circle their level of participation in creating or selecting forms, the results being indicated in (Table 5).

The study employed quantitative indicators of knowledge levels on the three knowledge domains of student course evaluations. Correlations were used to test the relationship between knowledge and attitude on the pre and post survey. Two separate correlations (one on the pre survey and one on the post survey) were run to determine if a stronger relationship would be found after the interventions. Mean scores on the pre and post survey were compared to test knowledge prior to, and after an intervention, and to test the impact of the three separate interventions. Differences
within group knowledge gains were tested to determine which of the three interventions had the highest impact. Likewise, attitude change was tested to determine if increases in knowledge would lead to a corresponding change in attitude.

The research question asked: to what extent does level of user knowledge of student course evaluation variables affect corresponding attitudes toward student course evaluation issues? A subsidiary question asked: to what extent can such attitudes be changed by educational interventions designed to increase knowledge? Another subsidiary question asked: how intensive an educational intervention was required to train users in the fundamental components of student course evaluations to significantly increase knowledge or improve attitudes? The research hypotheses tested were as follows:

At T1, subjects with higher knowledge will demonstrate significantly more favorable attitudes towards student course evaluations.

At T2, subjects with higher knowledge will demonstrate significantly more favorable attitudes toward student course evaluations.

At both T1 and T2, administrators will demonstrate significantly more knowledge and significantly more favorable attitudes toward student course evaluations.

Between T1 and T2, those subjects afforded more intensive educational experiences will demonstrate significantly higher knowledge and favorability (Workshop Group) than those subjects offered less intensive (Reading Group) or no education at all (Control Group). For the (Workshop Group), knowledge will increase significantly more than the (Reading Group) and will result in a significant increase in attitude.
For the Reading Group, knowledge will increase more moderately from T1 to T2 than
the (Workshop Group) but greater than the (Control Group).

For the (Control Group), knowledge will not significantly increase from T1 to T2 and
there will not be a significant change in attitude.

To determine level of knowledge, participant responses were given a
numerical value. A five-response licker scale was used to guide data entry. If a
respondent's choice on a question represented the "highest knowledge" response, data
entry for that question became a five. That would indicate that the respondent was
most familiar with the literature, as it related to each particular question. Responses
to knowledge questions were coded on a five point scale, from 5 = most
knowledgeable, 4 = knowledgeable, 3 = Uncertain, 2 = less knowledgeable 1 = not
knowledgeable. Accordingly, the five attitude questions were coded on a five point
scale from 5 = most favorable, 4 = favorable, 3 = Uncertain, 2 = less favorable 1 = not
favorable.

Quantitative Findings

Hypothesis 1 and Hypothesis 2

The first two hypotheses of this study centered on the relationship between
knowledge and attitude. It was hypothesized that subjects with higher knowledge
will demonstrate significantly more favorable attitudes toward student course
evaluations. In the pre survey, variable one was created on knowledge questions (6-
20) and variable two was created on attitude questions (21-25). A Pearson
Correlation Coefficient (Pearson r) was run to test the relationship between
knowledge and attitude. Although there is no definitive standard for interpreting size
of correlation, there are some guidelines that are widely recognized. According to Cohen (1977), a coefficient of .50 or higher is large; Guilford (1956, p. 145) suggests that an r-value between .40 and .70 indicates a "substantial relationship." The r-value on the pre survey between knowledge and attitude was .540, as indicated in (Table 6).

On the post survey, variable three was created on knowledge questions (6-20) and variable four was created on attitude questions (21-25). The r-value on the post survey between knowledge and attitude was .616, as indicated in (Table 7). In both cases the correlation was significant at .001 level. In both cases the correlation suggests that the greater the knowledge, the more positive the attitude. The stronger correlation between the post survey knowledge and attitude variables may also suggest an impact of one or more of the educational interventions. Therefore, hypotheses one and two supported the theory that higher knowledge will lead to higher favorability.

Table 6

The Relationship Between Knowledge and Attitude on the Pre Survey

<table>
<thead>
<tr>
<th></th>
<th>Pre Knowledge</th>
<th>Pre Attitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre Knowledge</td>
<td>1.000</td>
<td>.540*</td>
</tr>
<tr>
<td>Pre Attitude</td>
<td>.540*</td>
<td>1.000</td>
</tr>
</tbody>
</table>

(N=93)

* p < .001
Table 7
The Relationship Between Knowledge and Attitude on the Post Survey

<table>
<thead>
<tr>
<th></th>
<th>Pre Knowledge</th>
<th>Pre Attitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre Knowledge</td>
<td>1.000</td>
<td>.616*</td>
</tr>
<tr>
<td>Pre Attitude</td>
<td>.616*</td>
<td>1.000</td>
</tr>
</tbody>
</table>
* p < .001

Hypothesis 3

A third hypothesis of this study was that on both the pre and post survey administrators will demonstrate significantly more knowledge and significantly more favorable attitudes toward student course evaluations than faculty. Using survey item number three, independent "administrator" and "teacher" variables were created. Using pre knowledge, pre attitude, post knowledge and post attitude as dependent variables, an Analysis of Variance (ANOVA) was run to test for significant differences between administrators and faculty. No significant differences were found between faculty verses administrators in knowledge or attitude, on either the pre or post survey. The results, presented in (Table 8) suggest that, although administrators usually make personnel decisions in which student course evaluations are offered as evidence of teaching performance, they were not in this case any more knowledgeable than teachers. The results also suggest that administrators did not demonstrate any more favorable attitudes than teachers did and that administrator
knowledge and attitude did not increase enough after the educational interventions to show significance.

**Table 8**

**Knowledge and Attitude Differences (Administrators versus Faculty): ANOVA**

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pre Knowledge</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>1</td>
<td>.219</td>
<td>641</td>
</tr>
<tr>
<td>Within groups</td>
<td>91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Post Knowledge</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>1</td>
<td>1.572</td>
<td>.213</td>
</tr>
<tr>
<td>Within groups</td>
<td>91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pre Attitude</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>1</td>
<td>.359</td>
<td>.551</td>
</tr>
<tr>
<td>Within groups</td>
<td>91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Post Attitude</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>1</td>
<td>.187</td>
<td>.667</td>
</tr>
<tr>
<td>Within Groups</td>
<td>91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hypothesis 4

The title of this study is, "The effect of an educational intervention on faculty and administrator attitudes to student course evaluations." Thus, a primary purpose of the study was to test the impact of the three interventions on the three separate groups. Mean scores on knowledge and attitude were computed for each group, on both the pre and post survey. The descriptive results of those mean scores are presented in (Table 9). To test for significance, independent variables, consisting of the Workshop Group, Reading Group and Control Group were created and dependent variables of pre and post knowledge and attitude were also created. The results indicate that no significant differences were found between any of the groups on pre knowledge and pre attitude, as represented in (Table 10).

However, significant differences were found between the groups on post knowledge and post attitude. Using an ANOVA to test for differences between groups, both post knowledge and post attitude were significant at the .01 level. The results of this test are represented in (Table 11). To further test which groups had significant differences between them individually, a Least Significant Difference was run. The results of this test, as represented in (Table 12), indicate that there were significant differences in post knowledge and post attitude between the Workshop Group and the Control Group, and between the Reading Group and the Control Group. There were not, however, significant differences between the Workshop Group and the Reading Group. The results of hypothesis four suggest that, in comparison to the Control Group both the Workshop Group and the Reading Group
demonstrated significantly higher knowledge and significantly more favorable attitudes as a result of an educational intervention.

**Table 9**

**Impact of Educational Interventions on Knowledge and Attitude (Descriptives)**

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pre Knowledge</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workshop Group</td>
<td>48.8182</td>
<td>5.1262</td>
<td>33</td>
</tr>
<tr>
<td>Reading Group</td>
<td>49.0000</td>
<td>5.8546</td>
<td>30</td>
</tr>
<tr>
<td>Control Group</td>
<td>48.1667</td>
<td>5.3065</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>48.6667</td>
<td>5.3065</td>
<td>93</td>
</tr>
<tr>
<td><strong>Pre Attitude</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workshop Group</td>
<td>18.4545</td>
<td>2.3196</td>
<td>33</td>
</tr>
<tr>
<td>Reading Group</td>
<td>17.8667</td>
<td>2.8616</td>
<td>30</td>
</tr>
<tr>
<td>Control Group</td>
<td>17.3000</td>
<td>2.5751</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>17.8925</td>
<td>2.6021</td>
<td>93</td>
</tr>
<tr>
<td><strong>Post Knowledge</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workshop Group</td>
<td>53.3636</td>
<td>6.7353</td>
<td>33</td>
</tr>
<tr>
<td>Reading Group</td>
<td>51.8333</td>
<td>5.9947</td>
<td>30</td>
</tr>
<tr>
<td>Control Group</td>
<td>47.6000</td>
<td>6.0663</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>51.0108</td>
<td>6.6863</td>
<td>93</td>
</tr>
<tr>
<td><strong>Post Attitude</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workshop Group</td>
<td>19.6364</td>
<td>3.4170</td>
<td>33</td>
</tr>
<tr>
<td>Reading Group</td>
<td>19.3333</td>
<td>2.0229</td>
<td>30</td>
</tr>
<tr>
<td>Control Group</td>
<td>17.2667</td>
<td>2.9819</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>18.7742</td>
<td>3.0436</td>
<td>93</td>
</tr>
</tbody>
</table>
Table 10

Impact of Educational Interventions ANOVA: (Pre Survey)

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre Knowledge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2</td>
<td>.202</td>
<td>.817</td>
</tr>
<tr>
<td>Within Groups</td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre Attitude</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2</td>
<td>1.568</td>
<td>.214</td>
</tr>
<tr>
<td>Within Groups</td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11

Impact of Educational Interventions ANOVA: (Post Survey)

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post Knowledge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2</td>
<td>6.975</td>
<td>.002*</td>
</tr>
<tr>
<td>Within Groups</td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post Attitude</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2</td>
<td>6.124</td>
<td>.003*</td>
</tr>
<tr>
<td>Within Groups</td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * p < .01
Table 12

**Impact of Educational Interventions: Least Significant Difference**

<table>
<thead>
<tr>
<th>Post Knowledge</th>
<th>M</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workshop Group-Reading Group</td>
<td>1.530</td>
<td>.337</td>
</tr>
<tr>
<td>Workshop Group-Control Group</td>
<td>5.763</td>
<td>.000*</td>
</tr>
<tr>
<td>Reading Group-Control Group</td>
<td>4.233</td>
<td>.011*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Post Attitude</th>
<th>M</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workshop Group-Reading Group</td>
<td>.3030</td>
<td>.678</td>
</tr>
<tr>
<td>Workshop Group-Control Group</td>
<td>2.367</td>
<td>.002*</td>
</tr>
<tr>
<td>Reading Group-Control Group</td>
<td>2.006</td>
<td>.007*</td>
</tr>
</tbody>
</table>

*Note: M = Mean Difference; N= 93; * p < .01*

**Hypothesis 5 and 6**

It was hypothesized (#5) that for the Reading Group, knowledge will increase more moderately (than the Workshop Group) from the pre to the post survey and that higher knowledge will lead to significantly more favorability (attitude). It was also hypothesized (#6) that the Control Group would not increase in knowledge and that there would not be a significant change in attitude. These two hypotheses were essentially testing the change from pre to post knowledge and the change from pre to post attitude within each group. Hypothesis five and six also addressed the difference in knowledge and attitude change between the groups.
To test these two hypotheses, a Paired Samples t-test was initially used. To test for significance of knowledge within the Paired Samples test, the mean score on knowledge in the pre survey was subtracted from the mean score on knowledge in the post survey. To test for significance of attitude within the Paired Samples test, the mean score on attitude in the pre survey was subtracted from the mean score on knowledge in the post survey. The results of the Paired Sample test(s) are presented in Table 13 (Workshop Group), Table 14 (Reading Group) and Table 15 (Control Group).

The results of the Paired Samples Test indicate that a significant improvement in knowledge was found in the Workshop Group and the Reading Group. There was not a significant difference in knowledge change for the Control Group. As regards knowledge change, these results confirm hypotheses five and six and suggest significant increase in knowledge was attained through the Workshop and Reading interventions. The results of the Paired Samples Test also indicate that a significant improvement in attitude was found in the Workshop Group and the Reading Group. There was not a significant improvement in attitude change for the Control Group. The results of the Paired Sample Test on attitude are presented in (Tables 13, 14, and 15).
### Table 13

**Knowledge and Attitude Change Within Groups from Pre to Post Survey:**

**Workshop Group**

<table>
<thead>
<tr>
<th>Paired Samples of Workshop Group</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge Pairs</td>
<td>Mean Difference</td>
<td>t</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>Post Knowledge-Pre Knowledge</td>
<td>4.545</td>
<td>4.907</td>
<td>.000*</td>
<td></td>
</tr>
<tr>
<td>(N=93); * p &lt; .001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attitude Pairs</th>
<th>Mean Difference</th>
<th>t</th>
<th>p</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Post Attitude-Pre Attitude</td>
<td>1.181</td>
<td>2.027</td>
<td>.05**</td>
<td></td>
</tr>
<tr>
<td>(N=93); ** p &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 14

**Knowledge and Attitude Change Within Groups from Pre to Post Survey:**

**Reading Group**

<table>
<thead>
<tr>
<th>Paired Samples of Reading Group</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge Pairs</td>
<td>Mean Difference</td>
<td>t</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>Post Knowledge-Pre Knowledge</td>
<td>2.833</td>
<td>2.529</td>
<td>.017*</td>
<td></td>
</tr>
<tr>
<td>(N=93); * p &lt; .05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attitude Pairs</th>
<th>Mean Difference</th>
<th>t</th>
<th>p</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Post Attitude-Attitude</td>
<td>1.466</td>
<td>3.612</td>
<td>.001*</td>
<td></td>
</tr>
<tr>
<td>(N=93); * p &lt; .001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 15
Knowledge and Attitude Change Within Groups From Pre to Post Survey:

Control Group

<table>
<thead>
<tr>
<th>Paired Samples of Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge Pairs</td>
</tr>
<tr>
<td>Post Knowledge-Pre Knowledge</td>
</tr>
<tr>
<td>(N=93)</td>
</tr>
</tbody>
</table>

| Attitude Pairs | Mean Difference | t    | p    |
| Post Attitude-Attitude | -0.033 | .098 | .923 |
| (N=93) |

Further Examination of Hypotheses 5 and 6

Hypotheses five and six examined the change from pre to post knowledge and the change from pre attitude to post attitude. To clarify, hypothesis four tested the significance of change in knowledge and attitude between the pre and post survey. The tests for hypothesis four analyzed mean knowledge and attitude scores; these tests on the pre and post surveys were run independently from each other.

Hypotheses five and six are the testing the significance of the change from the pre to post survey.

The initial Paired Samples tested the significance of that change by analyzing each group separately. To test for the significance of change overall, or between the three groups, two tests were run. The first was an ANOVA, which tested the pre to post change in knowledge and the pre to post change in attitude, between the groups. The results, presented in (Table 16), suggest that the change in knowledge was
significant but the change in attitude was not significant. A second test, a test of Least Significant Differences, shows the significance of change in knowledge and attitude by individually comparing each group to the other two. The results of the Least Significance Difference test, presented in (Table 17), suggest that significant change in knowledge resulted when comparing the (Workshop Group) to the Control Group and also when comparing the (Reading Group) to the Control Group. The test further suggests, also in (Table 17), that the change in attitude was not significant when comparing the Workshop Group to the Reading or Control Groups. The test does indicate significant change in attitude when comparing the Reading Group to the Control Group, but that was not supported by the ANOVA when comparing all three.

**Table 16**

**Knowledge and Attitude Change Between Groups: ANOVA**

<table>
<thead>
<tr>
<th>Knowledge</th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre to Post Change</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In Knowledge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2</td>
<td>8.969</td>
<td>.000</td>
</tr>
<tr>
<td>Within Groups</td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attitude</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre to Post Change</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In Attitude</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2</td>
<td>2.861</td>
<td>.062</td>
</tr>
<tr>
<td>Within Groups</td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 17

Least Significant Difference Test of Change

<table>
<thead>
<tr>
<th>Pre to Post Change in Knowledge</th>
<th>M</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workshop Group-Reading Group</td>
<td>1.712</td>
<td>.165</td>
</tr>
<tr>
<td>Workshop Group-Control Group</td>
<td>5.112</td>
<td>.000*</td>
</tr>
<tr>
<td>Reading Group-Control Group</td>
<td>3.400</td>
<td>.008*</td>
</tr>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pre to Post Change in Attitude</th>
<th>M</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workshop Group-Reading Group</td>
<td>-.2848</td>
<td>.664</td>
</tr>
<tr>
<td>Workshop Group-Control Group</td>
<td>1.215</td>
<td>.066</td>
</tr>
<tr>
<td>Reading Group-Control Group</td>
<td>1.500</td>
<td>.027**</td>
</tr>
<tr>
<td>(N=93)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. M = Mean Difference; * p < .01; ** p < .05

Summary of Data Analysis

The impact of an educational intervention yielded a variety of results from the data analysis. Level of knowledge seemed to affect subjects' attitudes regarding student course evaluations. Knowledge and attitude variables were strongly correlated in the pre survey. Those same variables were also substantially correlated in the post survey. Thus, it is suggested that the higher the knowledge level, the more favorable the attitude. It is also a possibility that the workshop and reading interventions influenced the higher correlation in the post survey.
In comparing the level of knowledge and attitude between faculty and administrators, no significance differences were found. Results suggested that on both the pre survey and post survey, administrators were not more knowledgeable than faculty and also did not demonstrate more favorable attitudes. This is an interesting outcome when considering the faculty evaluation process, as that process regards student course evaluations. The literature points out that student course evaluations are almost always used in the faculty evaluation process. While it usually administrators that decide on retention, merit pay, promotion and tenure, they did not demonstrate any higher knowledge or more favorable attitudes. They also did not improve in either domain with any significance, as compared to the faculty.

The impact of the workshop and reading interventions did, however, show significance in the post survey. While no significant difference in knowledge or attitude was found in the pre survey, the post survey revealed significant differences between the Workshop Group and the Control Group, and between the Reading Group and the Control Group. This significance was demonstrated on knowledge and attitude. No significant difference, in knowledge or attitude was seen between the Workshop Group and the Reading Group. Thus, an intervention was able to raise the knowledge level of subjects and also improve their attitude toward student course evaluations.

The data also showed that significant difference was found between the pre survey and post survey on knowledge within the Workshop and Reading Groups. The Control Group did not demonstrate significant improvement from the pre survey to the post survey. The Workshop Group and the Reading Group also showed
significant improvement in attitude from the pre to post survey, while the Control Group did not. When comparing the change between the groups, knowledge change was significant. Attitude change between the three groups did not show significance in an ANOVA, but a Least Significant Difference test showed significance between the Reading and Control Groups. A summary of the results of this study is presented in (Table 18).

**Table 18**

**Summary of Data Analysis**

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Analysis</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1: Relationship between Knowledge and Attitude</td>
<td>Correlation</td>
<td>p &lt; .001</td>
</tr>
<tr>
<td>#2: Relationship between knowledge and attitude</td>
<td>Correlation</td>
<td>p &lt; .001</td>
</tr>
<tr>
<td>#3: Administrators will demonstrate significantly more knowledge than faculty</td>
<td>ANOVA</td>
<td>NS</td>
</tr>
<tr>
<td>#4: Administrators will demonstrate significantly more favorability than faculty</td>
<td>ANOVA</td>
<td>NS</td>
</tr>
<tr>
<td>#4: Impact of educational intervention on knowledge</td>
<td>ANOVA</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>#4: Impact of educational intervention on attitude</td>
<td>ANOVA</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>#5 &amp; #6: Knowledge change from pre to post survey within the Workshop Group</td>
<td>Paired Samples</td>
<td>p &lt; .001</td>
</tr>
<tr>
<td>#5 &amp; #6: Attitude change from pre to post survey within the Workshop Group</td>
<td>Paired Samples</td>
<td>p &lt; .05</td>
</tr>
<tr>
<td>#5 &amp; #6: Knowledge change from pre to post survey within the Reading Group</td>
<td>Paired Samples</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td></td>
<td>Method</td>
<td>Statistic</td>
</tr>
<tr>
<td>------------------------</td>
<td>---------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Attitude change</td>
<td>Paired Samples</td>
<td></td>
</tr>
<tr>
<td>from pre to post</td>
<td>within the Reading</td>
<td></td>
</tr>
<tr>
<td>survey</td>
<td>Group</td>
<td></td>
</tr>
<tr>
<td>Knowledge change</td>
<td>Paired Samples</td>
<td></td>
</tr>
<tr>
<td>from pre to post</td>
<td>within the Control</td>
<td></td>
</tr>
<tr>
<td>survey</td>
<td>Group</td>
<td></td>
</tr>
<tr>
<td>Attitude change</td>
<td>ANOVA</td>
<td></td>
</tr>
<tr>
<td>from pre to post</td>
<td>survey between the</td>
<td></td>
</tr>
<tr>
<td>between the three</td>
<td>three groups</td>
<td></td>
</tr>
<tr>
<td>groups</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge change</td>
<td>ANOVA</td>
<td></td>
</tr>
<tr>
<td>from pre to post</td>
<td>survey between the</td>
<td></td>
</tr>
<tr>
<td>between the three</td>
<td>three groups</td>
<td></td>
</tr>
</tbody>
</table>

Note. (N=93); (NS=Not Significant)
CHAPTER V

DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

Introduction

Few topics in higher education generate as much debate as student course evaluations. Stakeholders express a variety of views on them, from overwhelming support to calls for their total removal. Some see them as a valid measure of teaching effectiveness and a reliable component of faculty personnel decisions. Others hold outright disdain of their use even for formative teaching improvement. The polarity of beliefs surrounding student course evaluations can be found in a variety of sources within the literature. Typically, the most extensive and widely cited research studies endorse student course evaluations as a valid and reliable basis to evaluate teaching effectiveness if not student learning. However, commentaries, which try to cast doubt on their efficacy, are regularly found in such venues as "The Chronicle of Higher Education."

With the weight of research evidence supporting student course evaluations as a primary tool to evaluate teaching, their use has risen sharply the last twenty-five years. Increase in the use of student course evaluations has also been bolstered by a demand for increased accountability in higher education. Student course evaluations are widely accepted by administrators as a practical means to assess teaching. Additionally, student course evaluations are a stable component of faculty personnel
decisions. Despite research evidence and general acceptance of them by administrators, faculty often complain that using student course evaluations is an unfair measure of their professional performance. Objection to the use of student course evaluations range from, their not being valid, to their compromising the professor's academic freedom.

The vast majority of prior research and publication on student course evaluations has centered on this polarity of views. The literature contains a great deal of supporting data on the validity and reliability of student course evaluations. These are found within the multi-section, multi-method, meta-analytic and bias studies literature. As a result, most competent researchers agree with Marsh (1987) that, "ratings are multidimensional, reliable and stable (and) primarily a function of the instructor who teaches the course rather than the course that is taught" (p. 253). The literature also contains some rather faulty but very popular studies that claim to refute the validity and reliability of student course evaluations. Those same suspect studies are often cited within the opinion pieces against ratings.

A review of the literature within this study highlights the contradiction between research evidence that supports ratings and negative sentiment opposing their use. It has been hypothesized in this study that the source of this contradiction can be found within the application of the student course evaluation process. Essentially it is bad practice by users that cause problems, not the instruments themselves. Lack of familiarity with the literature and lack of cognitive development in the three knowledge domains can lead to bad practice and negative sentiment. The
fundamental knowledge components of student course evaluations and the association between knowledge level and corresponding attitude was a focus of this study.

This study sought to contribute to the emerging literature on student course evaluations in a few ways. First, it tested the relationship between knowledge and attitude. It was hypothesized that participants who had higher knowledge would demonstrate a more favorable attitude toward student course evaluations. The study then tested the impact of an educational intervention on both knowledge and attitude. Three groups were the focus of this study, two that received an intervention and a control group. It was hypothesized that participants who received a more intensive educational intervention would show significant improvement in knowledge and attitude; the Workshop Group demonstrating the most improvement and the Reading Group demonstrating more improvement than the Control Group. The study also tested if there were significant differences between administrator knowledge and attitude and faculty knowledge and attitude. It was hypothesized that administrators would demonstrate significantly more knowledge and significantly more favorable attitudes toward student course evaluations.

Therefore, the study sought to contribute to the emerging literature by first highlighting the relationship between knowledge and attitude. Franklin and Theall (1989) demonstrated this relationship, and this study sought to revisit what they had found twelve years earlier. It was hoped that if a substantial relationship was found, a recommendation could be made to increase user knowledge, as it might also increase favorability. The rationale was that individuals who are most knowledgeable could also be those individuals who express the most amount of favorability. Based on that
suggestion, the study also sought to demonstrate that an educational intervention could increase knowledge and attitude, and sought to suggest that educational interventions accompany the establishment of any ratings system.

Discussion of Results

There were six specific hypotheses that were tested in this study. The first two centered on the relationship between knowledge and attitude. They were stated: "subjects with higher knowledge will demonstrate significantly more favorable attitudes towards student course evaluations." To test hypotheses one and two, a Pearson Correlation was run. Variable one (knowledge on the pre survey) was matched with variable two (attitude on the pre survey), while variable three (knowledge on the post survey) was matched with variable four (attitude on the post survey). The results indicate a significant relationship between the knowledge and attitude. Statistically, a .54 correlation was found in the pre survey and a .61 correlation was found in the post survey, between knowledge and attitude.

These findings are significant for several reasons. First, it can be suggested from the research that higher knowledge does lead to more favorable attitudes. This is particularly important given the debate surrounding student course evaluations. It is important for higher education stakeholders to contemplate level of user knowledge when designing and implementing a ratings system. Second, it can be inferred from the results that the two educational interventions had an impact on both knowledge and attitude, as the correlation became stronger when the post survey was analyzed. Moreover, the mean knowledge and attitude scores increased from the pre to the post survey. Third, the correlation between knowledge and attitude in the post survey,
while substantial, also includes the Control Group. The Control Group did not increase in knowledge nor did they increase in favorability. Had they received an intervention similar to the Workshop or Reading Group the post survey relationship between knowledge and attitude might have been stronger. It would be beneficial to any ratings process to consider the possibility of a fundamental relationship between knowledge and attitude and to contemplate the potential improvement in both through educational interventions.

Hypothesis three tested administrator knowledge and attitude verses faculty knowledge and attitude. It was stated: "administrators will demonstrate significantly more knowledge and significantly more favorable attitudes toward student course evaluations." To test this hypothesis, an ANOVA was run to determine if significant differences in knowledge or attitude were present in either the pre or post survey. The mean knowledge and attitude score of administrators and faculty on both the pre and post survey did not reveal any significant differences between the two groups. The $p$ value on all four measures (pre knowledge, pre attitude, post knowledge, post attitude) did not approach significance.

The results present an interesting reality regarding the development and use of student course evaluations in personnel decisions. Clearly, administrators are at the forefront of faculty personnel decisions. As the literature suggests (Centra, 1979; Miller, 1987; Arreola, 1994) administrators need to educate and include stakeholders who will be affected by any evaluation process. For a system to be truly effective it is incumbent upon administrators to be the most knowledgeable and have the best attitude toward student course evaluations. While it hard for this study to generalize
to all of higher education, there were administrators with experience creating and selecting forms that participated in the study. That same group of administrators also had participated in personnel decisions in which student course evaluations were offered as evidence of teaching performance. While administrators in this study had created and selected forms, in addition to participating in personnel decisions, they were not any more knowledgeable than faculty, nor did they demonstrate any more favorable attitudes.

Hypothesis four tested the impact of the educational interventions on knowledge and attitudes. It was stated: "between T1 and T2 those subjects afforded more intensive educational experiences will demonstrate significantly higher knowledge and favorability (Workshop Group) than those subjects afforded less intensive (Reading Group) or no education at all (Control Group)." There were no significant differences in knowledge and attitude between the groups on the pre survey. The results of the ANOVA show significant differences in knowledge and attitude ($p < .01$) between the groups in both knowledge and attitude on the post survey. Specific examination of the mean score differences in knowledge and attitude with the Least Significant Difference (LSD) suggests that both the Workshop Group and Reading Group had significant differences between them and the Control Group in the post survey. The LSD did not show any significant difference between the Workshop and Reading Group. That is to say, the two groups that received an educational intervention showed significant differences as compared to the Control Group. The two intervention groups did not show any significant differences when compared to each other.
The significance found between the groups suggests that the educational interventions were a productive way to increase user knowledge and at the same time improve user favorability toward student course evaluations. The Control Group score on knowledge and attitude actually went down between the pre and post survey. It can be inferred that the time between the pre and post survey was not used by the Control Group to investigate the literature on student course evaluations or the knowledge domain items on the pre and post survey. It can also be inferred that the time between the pre and post survey, without intervention, did not help improve the attitude of the Control Group toward student course evaluations.

Demonstrating improvement in knowledge and attitude is significant because the topic of student course evaluations is so contentious. The mean knowledge scores on the pre survey were almost identical and did not show statistically significant differences among groups. Likewise, the mean attitude scores on the pre survey were very similar and also did not show statistically significant differences among groups. The movement in both knowledge and attitude on the post survey are an indication of the possible benefits of educational interventions. While the Workshop Group did not demonstrate significant difference in knowledge or attitude compared to the Reading Group, it is important to note that both interventions were designed to achieve the same end (improvement in knowledge and attitude). The fact that both approaches did not significantly differ in their impact suggested that any approach may be beneficial and that not providing an intervention is certainly not likely to improve knowledge or attitude.
Hypothesis five and six tested the change in knowledge and attitude within each group. Succinctly put, it was hypothesized that significant change in knowledge and attitude would occur in the Workshop and Reading Groups, and the change within the Reading Group would be "more moderate." For the Control Group there would not be a significant change in knowledge or attitude. For the Workshop Group there was significant increase in knowledge ($p < .01$) and attitude ($p < .05$). For the Reading Group both knowledge ($p < .05$) and attitude ($p < .01$) increased significantly. The Control Group did not demonstrate any significant increase in knowledge or attitude. Essentially "within group change" confirmed both hypotheses five and six. Analyzing the change (from pre to post survey) "between the three groups" demonstrated significant increases in knowledge, supported by the ANOVA test. The ANOVA did not, however, show significant difference in attitude "between the groups."

The results of testing hypotheses five and six provide support for the use of an intervention to enhance user knowledge and attitude toward student course evaluations. If user knowledge can be significantly increased with an intervention, it can be inferred that the knowledge levels of users coming into the study were low or at least in need of improvement. Given the breadth of research publications (particularly since Franklin and Theall 1989) aimed at improving evaluation practice, the results of this study suggest that there are still problems with user understanding of the field. Each group demonstrated a similar level of knowledge on the pre survey, suggesting a comparatively low level of knowledge prior to the study.
Another important consideration based on the results of hypotheses five and six concerns the change in attitude of the two intervention groups. It was hypothesized that the Workshop Group would demonstrate a significant increase in attitude and that the Reading Group would increase more moderately compared the Workshop Group. The results of the post survey indicate that the attitude of the Reading Group actually increased more than the Workshop Group, with no significant difference between the two. Both groups demonstrated improved attitude, and the fact that the Reading Group demonstrated more improvement suggests several interesting possibilities. One possible consideration is that the short-term intervention of the Workshop Group did not allow enough time for them to deliberate on the attitude items on the post survey. Conversely, the Reading Group had more time to reflect upon those attitude items, in addition to having reading materials in front of them. While the study did not produce tangible evidence to support those possibilities they are useful considerations in designing or evaluating educational interventions for a ratings system.

Conclusions

The results of this study indicate that it is possible to increase user knowledge of the fundamental knowledge domains of student course evaluations. Significant differences in knowledge were found between the three groups on the post survey. Additionally, significant increases in knowledge were found on the post survey in both the Workshop Group and the Reading Group. The fact that the Control Group decreased its mean score on the post survey offers evidence that an educational intervention is required to improve knowledge. Scores of all three groups on the pre
survey suggest that users are not as knowledgeable about student course evaluations as they could be. Ratings system designers can be confident that an intervention designed to increase user knowledge can be effective toward that end.

The results of this study also indicate that user attitude can be improved through an educational intervention. Both the Workshop Group and the Reading Group increased their favorability toward student course evaluations. To again use the Control Group for comparison purposes, their lower attitude score on the post survey is an indication that without an educational intervention, user attitude toward student course evaluation may not change. Lack of any intervention can be considered bad practice, which in turn can reinforce negative sentiment.

Another indication from the results of this study is the relationship between knowledge and attitude. Systems designers might be persuaded to implement an educational intervention, based on the substantial correlation found here. It was hypothesized that the Workshop Group would demonstrate the most improvement in attitude, it was the Reading Group that actually improved the most. One possible explanation for this is that the short-term intervention did not allow Workshop Group participants enough time to reflect upon their attitude responses. Dewey (1974) and Schon (1987) suggest that reflection upon a topic requires time for extended consideration, to move beyond the initial period of cognitive dissonance, toward actual change in belief and attitude. It may be inferred that any intervention will enhance knowledge and attitude.

Student course evaluations is a topic that higher education stakeholders are passionate about (often, with negative sentiment). For instance, Simpson and Siguaw
(2000) report from a sampling of faculty opinions that suspicion about the misuse of ratings can and has led to unethical attempts by faculty to use coercive methods to get high ratings. Considering that and concerns like it, is critical to educate stakeholders on the appropriate development of ratings systems. However, it may be hard for any institution to significantly increase favorability toward them if only a short-term intervention is used. While knowledge may increase in the short-term, attitude change on emotionally charged topics is much harder to accomplish. Rokeach (1960) defines such tenets as "anchor beliefs." What the results of this study suggest is that a best practice approach should include a sustained effort to increase user knowledge and attitude.

A concern stemming from the results of the study centers on hypothesis three. It was hypothesized that administrators would demonstrate significantly more knowledge and significantly more favorable attitudes toward student course evaluations. The results of the study indicate that neither was the case. The literature is marked with commentaries by higher education stakeholders about the improper use of student course evaluations, specifically by administrators who are responsible for their development and implementation, as well as for making personnel decisions based on those ratings. Trout (1997) cites poor administrative use of student course evaluations as a cause of lower standards and Taylor and Pierce (1999) talk about faculty "blaming their supervisor" for poor ratings. It is crucial for administrators to lead by example in the case of student course evaluations. In designing a ratings system, it should be administrators that receive the first educational interventions. Emanating from those initial sessions should be a common knowledge base and a
shared positive disposition toward the use of student course evaluations. Without a foundation for a best practice approach, established by administrators, most faculty evaluation systems will be less than productive.

This study has contributed to the literature in a few ways. First, the study in one sense revisited Franklin and Theall (1989) as it made the connection between knowledge and attitude. The relationship between knowledge and attitude is a fundamental consideration of a best practice approach to student course evaluations. The high correlation found in the study should serve as a central consideration for users. In another sense the study also confirmed Franklin and Theall (1999) who found a "lack of knowledge" by ratings users. In this study, the pre survey mean knowledge scores can be inferred as a lack of knowledge.

Improving that lack of knowledge was addressed through another contribution this study has made to the literature. That contribution came through the influence of the educational interventions on both knowledge and attitude. The Workshop Group experienced the greatest knowledge gain and the Reading Group experienced the greatest gain in attitude. The seemingly low scores on the pre survey were significantly improved in the post survey by way of the two interventions. Ratings systems designers need to contemplate the type of intervention they want to implement to develop a best practice approach.

One way to develop a best practice approach would be utilize the research design employed in this study. That design represents a third contribution to the literature made by this study. There are dozens of studies that tested the validity and reliability of student course evaluations, and other publications that have focused on
perceptions or opinions of stakeholders regarding student course evaluations.

Franklin and Theall (1989) have tested knowledge and attitude, employing a survey method. This study, however limited, represents the only research on student course evaluations that applied educational interventions designed specifically to increase user knowledge and improve user attitude. It also represents the only study on student course evaluations in which participants were asked to complete a pre and post survey as measures of knowledge and attitude. Combined with the three distinct groups that were the focus of the research, this study represents a unique contribution to the emerging literature on student course evaluations.

Recommendations for Further Research and Practice

The study's findings suggest opportunities for additional research in the field of faculty evaluation, as it specifically relates to student course evaluations. The study results also present a chance for higher education stakeholders to replicate the research design. The research results provide ratings systems designers the potential to increase user knowledge and improve user attitude. In an effort to establish a best practice approach toward the use of student course evaluations, including both formative and summative purposes, the following recommendations for future research are offered:

1. Analyze the three knowledge domains (obtaining, interpreting and communicating data) separately. Test to determine if there were significant differences between the three domains. It could be suggested, based on the results, that users are least knowledgeable in one particular domain or another.

For instance, if it is determined that knowledge of interpreting data is the domain
that users are least knowledgeable in, both on the pre and post survey, then a separate intervention that targets statistical knowledge, may be recommended.

2. Analyze if level of user participation in personnel decisions or participation in creating or selecting forms resulted in higher knowledge or more favorable attitudes. Based on the results of hypothesis three, it would be important to establish the knowledge and favorability levels of users who are most responsible for the design and implementation of ratings.

3. Recreate the research design with a truly random sample. One of the limitations of this study was the sample selection. An important follow-up research study would be to compare the three groups through a representative sample (i.e. and entire university faculty). Due to the challenging nature of subject recruitment for this study, the most feasible avenue to accomplish this would be through a campus-wide workshop. The opportunity for this is during staff development days held at the beginning of each academic year. Each participant would need to be assigned to one of the three groups by chance. Three hours would be enough to complete the Workshop, Reading and Control Groups.

4. Replicate the study to account for institutional type and discipline. It may that the results of this study are influenced by either (or both of these) variables. The small sample size of this study does not allow for analysis by institutional type or discipline. The literature on faculty suggests that institutional type and discipline influence faculty values, activities and attitudes. There may be certain disciplines (i.e. Education) or certain institutional types (i.e. community colleges) that have
greater knowledge or favorability to begin with. The implications to be drawn for those subgroups remain uncertain and require further research.

5. Conduct a follow-up survey of the Workshop Group and the Reading Group. Essentially, it is most important to determine if knowledge levels have been sustained and if favorability has been maintained. The interventions were successful in the short-term, but a larger question remains about long-term retention of knowledge and attitude levels.

6. Survey New Jersey colleges and universities to determine if any type of educational intervention regarding student course evaluations has taken place since March 30, 2001. The mean knowledge and attitude scores of the Control Group on the post survey suggest that no intervention will certainly not increase knowledge nor improve attitude. If the majority of colleges and universities have not implemented an educational intervention, it is possible that negative sentiment or even bad practice will be reinforced.

Accordingly, the following recommendations for future practice are offered:

1. Establish a standing administrative committee, whose purpose is to design, implement and oversee the faculty evaluation process. This committee must be representative of all campus departments or colleges. Within this group, one of the focus areas needs to be the development and use of student course evaluations. Once established, the committee should host an administrative workshop designed to educate participants in the fundamental knowledge domains of student course evaluations. This workshop should be facilitated by a leading expert(s) in the
field of student course evaluation research. The workshop should be taped and reproduced for viewing purposes around campus.

2. Once the workshop is completed, every faculty and administrator on campus should be required to complete the pre survey used in this study. Once all the pre surveys are completed, every faculty and administrator on campus should watch the video. Additionally, each faculty member and administrator should be given required reading materials that complement and enhance the topics covered on the video. Once completed, all participants should complete the post survey used in this study.

3. The pre and post surveys should be analyzed statistically and the results should be presented publicly to the entire campus. Each of the knowledge domains should be reviewed to provide a clear understanding.

4. The committee should develop a student course evaluation questionnaire that is two-fold. The first component should include items that every student in every course will complete. The second should include items that are specific to specific departments or colleges within the university. Thus, each department or college within the university will be able to address their unique curricular and evaluative concerns. A panel of experts external to the college or university should validate the construct of the questionnaires. Changes recommended by the panel should be made and again reviewed for validation. This recommendation is consistent with the literature, which suggests that ratings systems cannot reliably compare across all disciplines and departments.
5. Ultimately, the framework for using ratings in formative and summative decisions should be debated and agreed upon. Strategies for improving teaching (based on ratings results) must be established prior to using the new questionnaires. Review of the system must be continual.
References


Seton Hall University. (2001). Workshop on student course evaluations. Seton Hall University, Colloquium conducted at Seton Hall University, South Orange, NJ.


Appendices
Appendix A

Workshop Group Recruitment Letter
January 26, 2001

Dr. Paul Douillard  
Vice President for Academic Affairs  
Caldwell College  
Caldwell, NJ 07006

Dear Dr. Douillard:

I am writing with an invitation (see below); and to alert you to a new opportunity that we have established in the higher education doctoral program at Seton Hall University—a “track” that focuses on the preparation and professional development of college teachers. The purpose of the new track is to provide current and prospective college teachers with (1) an understanding of adult learning and development, group processes, instructional design and technology; (2) clinical opportunities to translate these understandings into their teaching; and (3) a broad understanding of current issues in American higher education that impinge on faculty work.

The new “College Teaching” track may be a useful vehicle for as yet undoctored faculty members who seek a terminal degree. Moreover, individual courses in the program may be of interest to individual faculty members as professional development opportunities. The new track complements the current Ed.D. Program in Higher Education Administration and the Ph.D. Program in Higher Education Research and Policy.

While it is primarily a degree-granting academic program, the new track also seeks to bring “cutting edge” thinking on college teaching to the attention of the broader higher education community through sponsorship of colloquia and conferences. The first such (inaugural) colloquium, which will be scheduled on Friday March 30 2001, from 12 noon to 3 PM will address a perennial issue in academic life: student course evaluations, the bane or boon of college teaching (depending on your perspective).

The Colloquium will feature Dr. Mike Theall, University of Illinois, a noted national expert on the psychometric characteristics of student course evaluations and a student of their effective and ineffective use on college campuses, and a panel of New Jersey faculty and administrative leaders. Dr Theall’s presentation will focus on “Student Course Evaluations: What Do They Measure, What are their Limitations and
What are their Valid and Invalid Uses." We want to invite up to 30 faculty, department chairs, deans and rank and tenure committee members statewide to attend. The cost is $10 per person, including lunch. Our only requirement is that participants complete a brief survey at the time they register for, and after completion of, the workshop.

Can you identify up to four faculty leaders, chairs, deans and promotion committee members from your campus who might be interested in, and benefit from, attending the Colloquium?

Enclosed is a flyer that you can distribute to your colleagues with basic information on the Colloquium. Can I ask you to circulate the flyer on campus and compile the names and contact information for interested individuals by February 5, 2001? I have asked Mr. Michael Rossi, a doctoral candidate in higher education, who will be coordinating the Colloquium, to telephone your office during the week of February 5th. Or, you may e-mail names and contact information directly to Mr. Rossi at: mrossi@centenarycollege.edu

I look forward to telling you more about the Ed.D. Program in College Teaching and to addressing with you some of the challenges of student course evaluations!

Best wishes,

Marty Finkelstein
Marty Finkelstein
Professor of Higher Education

CC: Mike Rossi
Joseph Stetar
Shouping Hu
Dean Richard Ognibene
Provost Mel Shay

Enc.: Flyer on Colloquium; Ed.D. Track in College Teaching Degree Requirements
Appendix B

Colloquium Registration Form
Inaugural Colloquium on
*Current Issues in College Teaching and Learning*
Presents

**STUDENT COURSE EVALUATIONS:**
What do they measure?
What are their limitations?
What are their valid and invalid uses?

**Date:** March 30th  
**Time:** 12-3 PM  
**Place:** Seton Hall University, Beck Room, First Floor (Walsh Library)  
**Cost:** $10 (Lunch Included)

Please complete the information asked for below and return, with ten dollars per person to:
Dr. Marty Finkelstein, Seton Hall University, 400 South Orange Ave.  
South Orange, NJ 07079-2697  
EDAS, 418 Kozlowski Hall.

Name of College or University: ____________________________________________

Mailing Address: ________________________________________________________
Please list each attendee’s name and email separately.

Name/Email ____________________________________________________________

Name/Email ____________________________________________________________

Name/Email ____________________________________________________________

Name/Email ____________________________________________________________

Name/Email ____________________________________________________________

Total number of attendees:  
@ $10 per person = ______

Make checks payable to Marty Finkelstein

*After completing this registration form, make copies of it for your own institutional purposes and mail the original, with payment to Dr. Finkelstein.
Appendix C

Reading Group Recruitment Letter
April 2, 2001

Dr. Paul Douillard
Vice President for Academic Affairs
Caldwell College
Caldwell, NJ 07006

Dear Dr. Douillard:

I am writing with an invitation to participate in a research project on student course evaluations. Few topics in higher education generate as much dialogue as student evaluations of teachers (SET). The study is designed to ask participants to respond to various questions regarding end of semester student course evaluations. Participants will be asked to complete a short survey and place it in a pre-paid envelope. They will then be asked to complete three short readings on SET and complete the survey again. Can you identify up to four faculty, chairs, deans and promotion committee members who might be interested in, and benefit from participating in this study?

Please forward this information to your colleagues with my name and information. Thank you, and feel free to contact me regarding this study.

Sincerely,

Michael A. Rossi, Jr.
Michael A. Rossi, Jr.
Centenary College
400 Jefferson Street
Hackettstown, NJ 07840
908.852.1400 x2117
908.813.1984 (fax)
mrossi@centenarycollege.edu
Appendix D

Reading Group Instructions
Dear Educator,
Thank you for agreeing to participate in this study. Please use the *guidelines below to complete the 2 surveys. In all, your participation will take just about 1 hour. Please contact me with any questions, and I would be happy to send you the research results of the entire study if you are interested.

Mike Rossi, Centenary College 908.852.1400x2117
mrossi@centenarycollege.edu

*

1. Complete the blue survey and place in mail (stamped envelope attached).

2. Read the 3 short articles AFTER blue survey is in mail.

3. Complete the gold survey and place in mail (stamped envelope enclosed).

THANK YOU!!!!!!!!!

PS. On the last page of both the blue and gold surveys, there is coding information that is asked for (year of birth & mother's maiden name). We want to ensure anonymity and also not make anyone feel uncomfortable. Thus, you may put any year and any maiden name on the surveys. For instance, you may say you were born in 256 BCE, and that your mother's maiden name was or is Abigail Adams. Feel free to fabricate! What is really important is that you write the same year of birth and the same maiden name on each form so we can code them accordingly.
Appendix E

Theall & Franklin (1990)
Student evaluations of instruction are often misunderstood or are met with direct hostility. Ratings "systems" are often very unsystematic indeed; at worst, they are seen as the punitive arm of administration. But ratings systems can provide vital information. How can their potential be realized?

Student Ratings in the Context of Complex Evaluation Systems

Michael Theall, Jennifer Franklin

Student "evaluations" are a corrupt practice of the '60s, one of the many from that era that I hope will be completely forgotten. They are an easy sop to the students from administrators... who are unwilling or unable to do anything to really improve teaching... I happen to think that there are some really rotten teachers... who should be forced to shape up or leave. But that's a job for strong deans and chairmen, not student "evaluators" and educationists like yourself.

---Anonymous quotation from faculty responses to a survey on student ratings, conducted by the authors

It seems that Kenneth Eble was right when he said (1983, p. 65), "No corner of the university... lacks faculty members who fulminate against student evaluations, with little or no examination of the large body of research... that underlies the practice."

But what about this "large body of research"? "What does it tell us? Marsh (1987, p. 255) succinctly summarizes student ratings as "1) multidimensional; 2) reliable and stable; 3) primarily a function of the instructor who teaches a course rather than the course that is taught; 4) relatively valid against a variety of indicators of effective teaching; 5) relatively unaffected by a variety of variables hypothesized as potential biases; and 6) seen to be useful by faculty... by students... and by administrators."
Marsh's conclusions agree with those found in other major reviews of the past twenty years (Centra, 1989; Costin, Greenough, and Menges, 1971; McKeachie, 1979). Past student-ratings research has concentrated largely on methodological and measurement issues and on the dissemination of this literature—a reasonable focus, given that the student-ratings process is indeed a measurement event. Some writers have offered guidelines (at various levels of detail) for the development of evaluation programs that include ratings (Braskamp, Brandenburg, and Ory, 1984; Centra, 1979; Doyle, 1983; Aleamoni, 1987; Miller, 1987; Seldin, 1984).

Thus, student ratings are seen to fill needs felt by students, administrators, and many faculty. The increased use of and reliance on student ratings in colleges and universities across the country (Seldin, 1989) is testament to the practicality of using ratings as a simple means of collecting evaluation data. But if there is consensus among researchers, and increasingly widespread acceptance of the practice of collecting ratings, why are campus debates on the issue frequently so acrimonious? Why are users of student-ratings data so often unaware of even the most fundamental precepts of the literature (Franklin and Theall, 1989)? What is wrong with ratings practice? Are student ratings falling short of the goals they were intended to serve?

Miller (1987) offers four characteristics of evaluation systems, which help to explain this phenomenon. He notes that such systems often evolve haphazardly, are sources of dissatisfaction by virtue of their very purpose, can never be expected to achieve full acceptance or provide complete satisfaction, and are legally questionable. Despite their ubiquity, it seems that ratings have not (and perhaps could not have) become a fully effective part of the everyday processes of personnel decision making and teaching improvement.

The weaknesses of ratings systems have little to do with the validity and reliability of ratings themselves. Ineffective use of ratings can very easily come about when policy is unclear, when there is no faith in the reliability of data-collection procedures, or when reports of results are disseminated without regard for how they may be used or by whom. Merely having valid and reliable student data in no way ensures that the information will be used appropriately or effectively.

As academics, many ratings participants have some familiarity with at least one aspect of ratings (such as the methodologies for survey questionnaires or data processing), and yet they lack a more comprehensive view, failing to see ratings systems as one part of evaluation systems, which in turn are one part of instructional systems. Further, these systems are fit into departments, colleges, and institutions. At each level, new pressures are brought to bear, and new requirements are imposed, because the purposes, standards, users, and uses of the system change. For example, at the college or institutional level, consider the problem of comparing the performance of faculty from different disciplines. As Cashin notes in Chapter Eight, differences across disciplines can affect ratings results; departments and institutions must decide how to deal with these differences when personnel decisions are made. Can a teacher of English composition be compared fairly to someone teaching a course in physics, even with classes of the same size at the same level? The answer is
maybe, but only if the comparison is on a general construct (such as overall teaching, effectiveness) and if very careful and complete analysis of institutional norms shows ratings in the two departments to be similar (with no significant differences between means on the item for classes of similar size, level, and nature), and if such comparisons are allowable under institutional policy, and especially if those who make the decision are aware of all the factors just enumerated.

Student ratings "systems" are made up of more than questionnaires, machine-scorable answer sheets, and computer-generated reports of results. Regardless of the qualifications of their users, ratings systems are complex aggregations of functional components and processes that act together to collect, analyze, report, or help users employ students' perceptions of the instruction they have received. Such aggregations may be chaotic and poorly articulated, or they may be "default" systems churning out incoherent noise. Conversely, they may be systematically planned and implemented to provide valid, useful information. They are never simple, although they are often treated simplistically.

Any organized effort to collect ratings requires development and implementation of processes that involve faculty, students, administrators, and institutional resources. Ratings processes operate in the contexts of policy, politics, and philosophy. They may affect faculty satisfaction and motivation and, ultimately, such instructional outcomes as student achievement, satisfaction, and retention. Whether ratings processes ultimately work to provide valid, reliable information about teaching performance depends both on the quality of the processes used to obtain the information and on the ability of receivers to use it appropriately. Ratings systems are used to tell people who have a serious investment in their profession how well they are performing in at least one part of their job. Unfortunately, these systems often make this sensitive information public, without regard to the quality of the information or the qualifications of those who will use it. The scope of bad practice is unknown, but it seems more likely when ratings systems are developed without a synoptic, coherent overview and when practice is not governed by a clear process derived from the accepted literature.

The consensus of evaluation writers and researchers (Theall and Franklin, 1989) is that literature in the field must expand beyond measurement questions and begin to see student-ratings issues and problems as complex configurations of events that are faced on a day-to-day basis by faculty, administrators, and practitioners. As Manicas and Secord (1983, p. 399) note, "Events are the conjunctures of structured processes and are always the outcome of complex causal configurations at the same, and at many different, levels." For evaluation researchers, the task is to conduct new kinds of investigations into questions about "complex causal configurations," and it means bringing past and present research into everyday practice, in clear and useful ways. It appears that "qualitative" (Fetterman, 1988) or "naturalistic" (Lincoln and Guba, 1985) methods would be very useful in investigating and understanding these configurations.

An understanding of ratings systems requires knowledge of general theory and specific knowledge of the variables involved, as well as their effects on and relationships to one another. The development of coherent ratings systems requires systematic descriptions, organizational entities, resources, and links among them in real instructional environments. Handbooks on evaluation (Millman, 1981; Doyle, 1983; Braskamp, Brandenburg, and Ory, 1984) have made great strides in identifying issues for practitioners, but more work remains to be done on conceptualizing the realm of practice. The more we as researchers know about the complexities of rating systems, the more we can develop the knowledge we need (in our other role, as practitioners) to recognize, link, and use all the parts of the systems in our practice.

Our immediate goal is to facilitate this new direction for inquiry and action, by describing student ratings as systems in the context of larger systems and by
regarding research and practice as necessarily concurrent and complementary activities. We view ratings in terms of the goals they serve, their functional components, and the conditions that influence or determine the outcomes they achieve. To that end, this chapter offers some schemata for classifying the components and processes of ratings systems, as well as the settings in which they operate. We are not doing this purely as an excuse in model building; it is crucial to good evaluation practice that we know and account for interactions and relationships within the system. The model we present here is intended to illustrate these multidimensional economically and graphically.

Where does a systematic evaluation process begin? It starts with analysis of the context and conditions of the situation, the needs of the system’s users, and the role that the evaluation will play. There are certainly precedents to this approach. Scriven (1967) has noted the importance of the role intended for any evaluation, particularly with respect to whether the evaluation is “formative” or “summative.” The role determines, to some extent, whether the information collected will be concerned with “instrumental” (process) or “consequential” (outcome) data; in any case, however, evaluation’s primary goal is always the same: an estimation of merit, worth, or value. Stake (1967) proposes a “judgment matrix” that incorporates “antecedents,” “transactions,” and “outcomes” in arriving at judgements based on specified “standards.” Stufflebeam’s (1968) “CIPP” model stresses context, input, process, and product as the categories of evaluative concern. Whether we choose “antecedents” or “context,” the starting point is the situation(s) in which evaluation takes place.

In his decision of the effects of context and environment on teacher evaluation in public school, McKenna (1981) offers an ecological paradigm: What “ecologies” are of concern in this discussion? At the broadest level, the influence of the external context must be acknowledged. What are some of the influencing factors? A partial list might include local and national government, the general economy and the particular condition of the local economy, politics, current and potential funding, social attitudes, and other factors. For example, Apps (1988) describes how the needs and demands of society can shape (and be shaped by) higher education.

Another so-called ecological system encompasses the institution and its environment. In the case of multisite state systems, another layer of context is added. Geis (1984), in his discussion of the context of evaluation, refers to the way institutions have changed from “monastery to industry.” The history, nature, direction, type, reputation, and location of an institution will affect its policies and, thus, its evaluation systems. The target clients of the institutions (intended students, as well as the world they will enter after graduation) will influence policy and direction. With respect to evaluation policy and procedures, the words and actions of the institution’s leaders will have a profound effect. If the message from leaders is that promotion and tenure are based on scholarship, then the nature, uses, and effects of evaluation will not be the same as at an institution that puts teaching first. This does not necessarily mean that evaluation and ratings systems will be better or more widely accepted at the teaching-oriented institution. In fact, when ratings from the basis of personnel decisions, the chances for opposition to ratings are perhaps even greater. The presence and status of faculty organizations or unions will also affect policy, particularly concerning the extent to which that policy is negotiable, public, and part of a formal contract. Birnbaum’s (1988) discussion of the “cybernetics” of academic organizations offers very useful insights into how different kinds of institutions operate, and his four models of institutional types (collegial, bureaucratic, political, and organized anarchy) clearly differ in their approach to and implementations of personnel decision-making procedures.

Within institutions, college and departmental differences represent the final level of environmental context and layers of complexity. The same categories of
impact are at work here as at the institutional level. Consider, for example, the often heard comparison that says colleges of nursing emphasize teaching and interpersonal interaction with students, while colleges of engineering deliberately make their courses difficult and offer little help to students, on the theory that they have to weed out those students who "can't hack it." We have not encountered research that substantiates or refutes this proposition; such investigation would certainly shed light on the ways in which different departments view, treat, and use student-ratings data. We can, however, report an anecdote from first-hand experience.

A faculty member publicly proposed that student ratings should be weighed so that "A" students' ratings would be worth four points, "B" students' ratings would be worth three points, and so on, with no value assigned to failing students' ratings. This person would be surprised to read Theall, Franklin, and Ludlow's (1990) report that "A" and "B" students provided 69.7 percent (from a sample of 2,381 courses, with 47,732 individual respondents) of the "worst" ratings on an "overall instructor" item while "D" and "F" students provided only 11 percent of such ratings. The distribution of grades for the entire sample was very similar to this subset; thus no conclusion could be drawn that poor students attempt to punish their instructors or that good students always rate their teachers highly. It seems that the instructor's (or department's) perception of the student rater may have as much to do with ratings interpretation and use as does the student's perception of the instructor.

Consider, too, the political problems of small departments, where the intimacies of personnel decisions are shared by everyone. What about the large department that offers required service courses and uses temporary instructors to teach them? Should evaluation policy be exactly the same for these teachers as for permanent staff? One generally accepted rule for faculty evaluation (see Arreola and Aleamoni, Chapter Three) is that multiple sources of data and/or several administrators of a student-ratings instrument are necessary for fair and methodologically correct evaluation. Does this rule apply here? Does the department have a responsibility to its students to review temporary instructors each term and avoid rehiring those who are poorly rated? Can ratings be legitimately used for this purpose? Each situation requires something different. How can we decide what evaluation process will be most effective?

**Evaluation of Instruction in Context: An Information Matrix**

We can begin by locating rating systems within faculty evaluation settings in higher education. To portray the breadth and interrelationships of evaluation's concerns, we have devised a matrix of information needs for evaluation, predicated on the purposes of the evaluation information. Figure 1 displays this realm of evaluation activity in three dimensions. Note that ratings are only one source of information in the model. Doyle (1983) offers such matrices ("cubic schemata") to structure the focuses, sources, and ways of transmitting information, but this matrix is oriented more toward assisting in decisions about the nature and extent of the information collected and about the people to whom it is reported and how reporting is done. From this general model, we will scale our discussion down to student-ratings systems and their uses. First, however, we offer an explanation of the overall matrix.
Figure 1. Evaluation Information Matrix

Purpose of Evaluation

At the left of the figure, some general purposes of evaluation are presented hierarchically with respect to the amount of information each purpose requires. Under the right conditions, data collected for a single purpose may be usable for multiple purposes. For example, some of the information collected for teaching improvement may also be used for decisions on promotion and tenure, as well as for inclusion in a catalogue of evaluation results for students' use in course selection. Even though the information is the same, however, the ways in which it is presented in these situations will be different (Theall, Franklin, and Birdsall, 1987). Another important issue is that when an effective system for evaluation is in place, it is inherently able to couple data-collection routines and analysis for teacher and course evaluation, on the one hand, with investigations for separate purposes, on the other. Later in this chapter, we will present examples of the productive joining of evaluation data with other data for research in several different areas.

Users of Information

On the right side of the figure are the users of evaluation results. Although students frequently provide evaluative information, it is important to remember they are evaluators only when the purpose of evaluation is course selection. In fact, the real evaluator is effectively the person who makes a judgment that is based on the data at hand. Thus, when student ratings of instruction are used for decisions about
promotion, tenure, and merit, promotion and tenure committees, department chairs, or deans make the judgements and are the real evaluators. Such decisions carry great responsibility, as well as the implication that those who make such decisions should be knowledgeable, valid, reliable judges. Recent research (Franklin and Theall, 1989) has shown that many who make personnel decisions based at least partly on student ratings, are unfamiliar with the most accepted and important literature in the field. This finding underscores the need for extreme caution in collecting, analyzing, and reporting data that may be used for personnel decision making, and it reinforces the importance of college and department systems and policies that protect both the individual and the institution.

**Sources of Information**

At the base of the figure are depicted common sources of information used when courses, instructors, or instructional events and programs are evaluated. These categories are generally self-explanatory, with the exception of the boxes marked "others" and "records."

"Others" means any valid observers or reporters and can include teaching-improvement experts or evaluators, faculty from other disciplines, or persons naive to the subject matter. The point to remember is that the role of the evaluation will guide choices made in this area. In personnel evaluation, the observer clearly must have the expertise to judge competence or quality. A third party brought in to judge, for example, the completeness and currency of course content must be knowledgeable if his or her comments are to be included in the dossier. For formative purposes, a range of expertise on the part of observers may be more valuable because naive observers may encounter difficulties similar to those experienced by students, but comments from such observers do not really belong in a dossier for promotion and tenure. Shore and others (1986) give several examples of what to include in a dossier and suggest useful guidelines for assembling and presenting such materials for decisions about promotion and tenure.

"Records" is a category that includes a number of items: the registrar’s list of course grades, data about course enrollments, general advising, direction of dissertations and theses, alumni reports, publications, citation counts, and service to the university, the profession, and the community. The important point in gathering these data is to be aware of the emphases or the institution and the weight assigned to each type of activity. Remember also that the "records" category includes the number, type, forum, and quality or papers, presentations, and publications, as well as successful grant applications, research projects, and professional activities. In many cases, this is a most important set of numbers.

Seldin (1989) reports on the factors that colleges consider in evaluating overall faculty
performance, as well as on the sources of information that colleges use to evaluate teaching performance. Four of Seldin's findings are noteworthy. First, the department chair remains the predominant source of information in faculty evaluation. Second, systematic use of student ratings has risen significantly, so that ratings have become the second most frequent source of data for teaching evaluation (only a fraction of a point behind the departmental chairperson's input). Third, classroom visits by colleagues have gained significantly, despite the many problems associated with the nonstandardized procedures often used in such visits. Fourth, scholarship, publications, and professional activities have increased in importance, although their position relative to other factors has remained the same. The implications of Seldin's study concern the importance of using multiple sources of data and of understanding the relative weights assigned to these sources by the institution, the college, and the department.

Synergism: A Useful Concept

Synergism is the joint action of discrete agencies, action whose total effect is greater than the sum of the agencies' effects when the agencies are acting independently. Synergism is not unidirectional, however; the combined effects of evaluative agencies can produce negative and damaging results, as well as positive and productive ones. Scenarios for two evaluation systems are described here to demonstrate these differences.

A Negative Synergistic System

The Context. A large research-oriented university, where emphasis on teaching varies by college and department, seeks a means of obtaining ratings data. There are two underlying reasons: a strong demand from students for a campuswide system of evaluation and publication of results (the students are politically active and quite insistent), and a need to bring evaluation results efficiently from a campuswide system into personnel decision making (this need is articulated only to the extent that this use of evaluative data is said to be planned for some point in the future and must be inexpensive). There are also rumors of faculty unionization; strong feelings on the issue have already surfaced.

The institution has no organized resources (that is, no agency for evaluation or teaching improvement) and no professional staff who are experienced or trained in developing or operating evaluation systems. The student newspaper regularly prints stories about the need for an evaluation system and about how student leaders ("us") have struggled against a slow and recalcitrant administration ("them") and against certain faculty who oppose evaluation (another "them"). Faculty are divided in their opinions, but even those who support student ratings are concerned that faculty ("us") will be targets of certain students who have trivial or venal agendas ("them"). Worse, faculty are also concerned that the administration (another "them") will use results capriciously or only for punitive purposes. The administration ("us") feels that its intent is honorable, that some students ("them") are being too demanding, and that some faculty (another "them") are subverting attempts to create a reasonable system.

Negative synergism is already at work because the energies of the participants are at least partly turned toward self-protection, rather than toward a unified approach to solving a problem.

In time-honored fashion, a committee of faculty, students, and administrators is formed to review the issues and develop a proposal. Other committees representing the three groups are also created informally, and certain individuals take it upon
themselves to investigate the issue. More negative synergism crops up here because the newly formed agencies are invested with different levels of authority and are reporting to different audiences. All these committees take the same path: to devise a questionnaire. Now there is even more negative synergism because each committee's investment in its own form will create dissension. Nevertheless, different forms are proposed; debate over questions ensues, and a compromise is reached on a questionnaire. Despite the questionnaire's intuitive appeal (based on items that are not so different from those seen on the sample forms used by the various committees), not all sides are pleased. More potential negativity appears at this point because dissatisfaction will lead to resistance to and subversion of the survey process.

The questionnaire includes items on overall rating of instructor and course, desire to take other courses from the same instructor, workload and difficulty, the instructor's "friendliness," the instructor's speaking ability and organization, the instructor's skill in using the blackboard, the instructor's fluency in English, and the instructor's punctuality, as well as items on readings, assignments, and tests. There is the potential for still more negativity here because all the items are considered equally important, all will be used for each intended purpose, and some are either technically incorrect or inappropriate for one or more of their intended uses. The items may be important in their own right (as in the case of fluency in English and its importance to students), but they are not useful for other purposes. How will instructors who are native speakers of English be compared to one another, or to non-native speakers? From one point of view, this question may even be discriminatory.

It is agreed that student government will be responsible for conducting the evaluations, that the computing center will process the forms and produce reports (one for the instructor, one for the department chair, one compiled report for publication), and that a faculty committee will oversee the entire process. More negative synergism occurs here because three separate groups are involved in the process, and no one administrator is really responsible for the quality or control of the system.

The Result. There is much hostility and suspicion in this system, particularly since reports are being forwarded directly to the administration. The system is inaccurate. It is often late in producing its reports, and most of those involved dislike it. Fewer and fewer courses get evaluated, because the turnover in student government makes it impossible to standardize data-collection procedures or train students in data collection. Long-standing departmental processes (some of which may be superior to the new system) either duplicate the effort or are disrupted. Several errors in operations lead to data from some instructors' courses being identified incorrectly as data from the courses of other instructors. The catalogue of results contains mean scores of the items, with no norms or other statistics for interpretation. Results get misused or misinterpreted. Finally, a lawsuit results because a professor is denied tenure on the basis of published results. The university loses the suit. Within three years, the system is dead.

This system's problems could not be attributed to anyone person or any one group. Much effort was expended by many people, whose intentions were positive and honorable, but there was no systematic planning. The complexity of the problem, the initial misdirection, and the interactions of the agencies at work led to a series of mistakes in planning and execution. The agencies involved did not focus their efforts, and the result was conflict. The institution became "a house divided," and the demise of the system was predictable.
A Positive Synergistic System

The Context. A similar institution, facing similar demands, takes a broad view and begins its efforts by assessing and clarifying the needs of all interested groups. A unified committee is formed, but its tasks are clearly defined as twofold: to use external experts and identify the important steps that must be taken in developing and implementing a system, and to propose a set of goals that the system can profitably serve. When these tasks are completed, consensus is sought about the importance, relevance, and achievability of the goals. Finally, when a few clearly important goals are established, discussion about the system itself begins (with external experts recommended again, to avoid common technical, methodological, practical, administrative, and other errors). (One of the most successful efforts of this kind was at Miami-Dade Community College, where planning and consensus took place over a three-year period; see Jenrette, 1989.)

The institution defines its first goal as supporting and improving teaching and responding to students' needs for a voice. To this end, the following list of system requirements was developed:

1. Provide summary data for students' use in course selection, in the form of a published catalogue of result
2. Provide detailed data for confidential teaching improvement and, to the extent possible, provide assistance in gathering, interpreting, and using these data for improving teaching
3. Provide data for personnel decision making, including reports and institutional, college, departmental, or other appropriate norms
4. Provide data for institutional use, in the form of compiled results for colleges or departments and for faculty or student groups

In effect, then, this should be a multipurpose system able to meet these various needs efficiently. This goal requires a cooperative effort on the part of several university agencies. For example, the registrar provides accurate information on courses and instructors, so that response sheets can be preprinted in the correct numbers and with correct information. The computing center provides support for data management and processing. The campus mail system gets ready for two or three large-scale distributions of materials each term. Departmental secretaries are aware of the system's needs for accurate information on courses and instructors, and they assist the evaluation agency by providing information about changes, dropped courses, and so on. The synergism of these agencies allows more accurate and timely operations than would be possible if individual units performed all tasks independently.

One element crucial to the success of the system was the determination that, despite the need for a cooperative effort, one unit should ultimately be responsible for the system's operation. It would be staffed by highly qualified, experienced individuals. There would be ongoing involvement of faculty, students, and administrators in this unit, to make the best use of the system and determine its future uses, but any discussions about the system and its uses would be guided by the expertise of this unit's staff. In other words, the process would be governed by the same systematic rules for planning and design that governed the development of the system itself. Details of operations and decisions about the kinds and formats of reports would be based on standards of practice drawn from the literature and on the specific needs of the institution. Clear policies for uses of data would be determined before any data were collected. Safeguards would be included to protect faculty, students, and the institution from errors and from misuse or misinterpretation of data. Finally, the system would undergo a field test for at least one year. During this period, no personnel decisions would be based on system reports. At year's end, decisions
about further use of the system would await full reports on its administrative operation and on the validity and reliability of instruments, the development of norms, and the reactions of the system's clients.

**The Result.** This system has a better chance of success. No system is perfect, but the design of this one is sensitive to users' needs and has built-in flexibility and contingency plans. For example, let us assume that nowhere on campus is there a fully current and correct file of faculty's addresses. This means that some materials will not reach some faculty. Is time built in to the schedule of evaluation events to accommodate last-minute requests for materials? Is there a quality-control mechanism that prevents duplicate questionnaires from being used? Is there a way to detect whether Professor X mistakenly used Professor Y's materials because they were delivered to the wrong room? Is there a way to improve the accuracy of addresses or enlist departments in quality control of distribution once materials reach departmental offices?

The prognosis for this system is that it will meet the direct needs of its users and serve the institution well. Are there additional benefits? The following discussion demonstrates that there are many potential uses and benefits of a well-articulated, multipurpose system.

**Synergism Within the System**

Given the contexts of evaluation and the interrelationships of variables, what actually happens in a ratings system, and how would this system for evaluation work to accomplish various tasks? Figures 2 and 3 outline the tasks of such a system, as well as its theoretical foundations.

In Figure 2, the activities of evaluation and improvement are shown as threefold: collection and analysis of evaluation data; interpretation or reporting of data by practitioners, including any resulting prescriptions for improvement; and actual provision of references, materials, consultation, and resources to clients.

**Figure 2. Activities of Evaluation and Teaching Improvement**

- Data collection and analysis
- References and resources
- Interpretation
- Prescription

**Figure 3: Realms of Theory and Practice**

measurement
Figure 4: Multipurpose System for Evaluation, Improvement, and Research
Figure 3 displays the realms of theory and practice that inform evaluation and improvement. Measurement and evaluation theory inform data collection and analysis. Theories of planned change (for example, those from communication, organizational development, and the diffusion and adoption of innovations) guide the ways in which data are interpreted and reported to clients. Theories of teaching and learning provide references and guide the development or use of resources.

Theory and practice do not operate independently, however. In a functioning system, there is constant interaction between the two. Figure 4 shows the interrelationships of research and theory with practice. The relationships between theory and practice are synergistic—that is, the activities of one area support and nourish those of the other. The multipurpose system for evaluation, improvement, and research depicted in Figure 4 not only offers direct evaluation and improvement services but also can develop knowledge about the realms of theory by using the data and experience gathered in practice. For example, evaluation can be used for research into student attitudes. (Abbott and others, 1989), course improvement (Cronbach, 1963), curriculum design and review (Diamond, 1989), assessment (Gray, 1989), pedagogical research (Murray, 1985), research into teaching and learning (Stark and Mets, 1988; Theall, 1986), and psychological research (Yarbrough, 1989). These variables and the very nature of their interrelationships become targets for new directions in evaluation research. Just as the synergism of agencies within the institution can produce unexpected benefits, so can the synergism of research and practice produce valuable results when data collected and processed for one purpose complement data used for other purposes.

This system can help with institutional research (for example, analysis of evaluations and performance of beginning or "at risk" students). It can provide information for such teaching-improvement activities as the training of teaching assistants (Nyquist, Abbott, and Wulff, 1989), based on analysis of evaluations gathered from courses taught by teaching assistants in various subjects. The system can also involve faculty in research or development projects supported by internal or external funding (for example, such a system has been used at Northeastern University to identify faculty exceptionally skilled at using class time, who then helped to develop a teaching-improvement unit on this topic; see Theall, Sorcinelli, Sorcinelli, and Franklin, 1989).

The positive interactions of the agencies in our example create more benefits than the originally envisioned system was seen to contain. This is possible not only because the operations and services of the system are defined by the context and by the needs of the system's clients but also because the system is run as a coherent whole by persons with appropriate expertise and authority with overall responsibility for operations. Institutional commitment and communication are other essential elements of success.

Conclusion

Teaching is a multidimensional activity, and student ratings reflect the variety and range of teaching behaviors, as well as the successes or failures of those who practice its art, craft, and science. Nevertheless, while researchers, practitioners, and clients of ratings systems all acknowledge the multidimensionality of ratings, these parties have not brought the multifunctionality of ratings to bear on other issues of importance to faculty, students, and institutions. From inside an efficiently operating student-ratings system, we can conduct research and provide services to the users of evaluation, as well as to those who supply data for the system's purposes. An understanding of the contexts of evaluation is crucial, as is the concomitant ability to identify how the intended uses and audiences of evaluation affect the kinds of data...
collected and the ways in which results are reported. To use a ratings system without appropriate concern for the protection of faculty, students, and the institution is to misuse the system and defeat its purpose.

The consensus of evaluation writers and researchers is that the psychometrics of ratings are no longer the most important focus of research. In fact, it is the interrelationship—the synergy—of research and practice that deserves our attention. These new directions in evaluation are centered on better understanding of teaching itself, of the variables at work when teaching or evaluation take place, and of the ways in which evaluation results are presented and used. We no longer look only for variables that may bias evaluation; we also look for ways in which research and practice may inform each other, help us answer questions about teaching and learning, and lead to better-informed decisions about people, programs, and performance.

References


Diamond, R.M. *Designing and Improving Courses and Curricula in Higher Education: A Systematic*


Theall, M. (exec. Producer), Sorcinelli, G. (producer, author) Sorcinelli, M. D. (producer, author), and Franklin, J. (instruct. designer). Effective Use of Time in the College Classroom. (Videotape and


*Michael Theall is associate professor and director of the Center for Teaching and Learning at the School of Education, University of Alabama, Birmingham.*

*Jennifer Franklin is senior research associate at the Center of Applied Social Research and manager of the Office of Instructional Evaluation at Northeastern University.*
Appendix F

Franklin & Theall (1990)
Communicating Student Ratings to Decision

Makers: Design for Good Practice

Jennifer Franklin, Michael Theall

Anyone who uses student ratings of instruction as a measure of teaching performance or instructional quality should be aware of standards for good practice and other crucial issues associated with ratings use. Faculty should be able use ratings effectively as an assessment of the quality of the instruction they offer and as protection against the misuse of their own data. Administrators should be able to use ratings fairly and efficiently in performance appraisal without exposing themselves and their institutions to liability for misuse. We have become increasingly concerned that too many of those involved in applying ratings to the full range of evaluation activities, from questionnaire construction to decisions on promotion and tenure and consultation for teaching improvement, may not be aware enough of important ratings issues. We are convinced that providing safe-guards against misuse of ratings data is an important task for those who plan or administer systems for collecting and disseminating student ratings of instruction. Our purpose in this chapter is dual: to discuss the problem of designing and implementing a critical component of ratings use—that is, methods for communicating ratings to those who use them—and, along the way, to demonstrate the usefulness of a systematic approach to identifying strategies or safeguards that can help prevent improper use of ratings results.

Over the past six years, we have been immersed in a full range of activities stemming from the establishment of a universitywide system for collecting student ratings of instruction. We designed, developed, implemented, routinely operated, and periodically evaluated a system for collecting ratings data and disseminating information to users. During this period, we collected ratings data from nearly 20,000 course sections. Providing consultation on evaluation-system development to other institutions gave us opportunities to explore similar activities in other settings. In the process of validating our instruments and the services we provide, we have conducted
research probing such diverse issues of validity and reliability as relationships between the locus of students' performance attributions and important ratings variables (Theall, Franklin, and Ludlow, 1990); the knowledge, attitudes, and practices of ratings users in a variety of higher education settings (Franklin and Theall, 1989); and the relationship between instructors' grading practices and ratings variables.

From instructional systems technology we brought some praxiological habits, including the use of a common analytical method for developing instructional systems (see Romiszowski, 1981; Dick and Carey, 1985). Adapting this approach to the problem of constructing a large-scale processing and service system, we defined a strategy for development that required us to (1) identify and understand the needs underlying the development process, (2) define goals (desired outcomes) for the system, (3) analyze the information and skill requirement of rating-system participants, including staff and clients, (4) construct and operationally describe tasks performed by staff and clients in terms of the conditions and events instrumental to achieving the goals, (5) evaluate the performance of the system and its components and revise as necessary, and (6) ultimately evaluate the outcomes of the systems, to characterize its effectiveness with respect to achieving its goals. We chose this approach, and we continue to apply it to ongoing operations, so that we can consider our rating system as a whole, as well as account for each of its functional components.

Definitions
This chapter concerns the problems we encountered in the development process. Before we explore the particulars, however, we pause here to discuss the relationship between the process of communicating ratings to users and the question of how ratings are ultimately used.

Student ratings of instruction can be used to inform research and practice in a wide variety of instructional settings and applications. The many possible combinations of purposes, types of users, sources of information, and aspects of instructional quality demonstrate that rating systems are complex entities, even when they are small-scale operations.

Rating systems can be defined as more or less organized aggregations of process intended to collect and disseminate student-ratings data, in order to provide information about instructional quality (especially teaching performance and course design) to users, who may include faculty, administrators, researchers in faculty development, evaluation, or institutional topics, and/or students, and who use inference or hypotheses derived wholly or in part from ratings data to inform decisions, choices, or plans for action that are intended to maintain or improve the quality of instruction provided or received. This definition may be compulsively descriptive, but we hope it has the virtue of covering the who, what, where, and why in a field where writers and readers alike have frequently focused on the construction and validation of ratings questionnaires to the exclusion of other essential components.

For purposes of discussion in this chapter, whoever plans a system is at least a de facto system designer; those who are routinely responsible for providing ratings to
others are system operators. One person may assume more than one of these roles, or even all of them. Rating systems vary in scope, from a single user conducting an informal evaluation of a class to a full-scale, campuswide, or even national ratings-data processing services.

Designing and Revising a Local Campuswide System

When we began our development project, we read much of the literature on ratings as we could obtain. We found plenty of practical information about constructing, administering, and validating student-ratings questionnaires (Berk, 1979; Marsh, 1987; Weimer, Parrett, and Kerns, 1988; Scriven, 1989). We also found discussions of how ratings fit into larger evaluation processes, such as personnel decision making and teaching-improvement consultation (Centra, Froh, Gray, and Lambert, 1987; Erickson and Erickson, 1979). These sources tend to emphasize contexts for ratings use. For example, Miller (1987) notes that ratings used for personnel decision making should always be used in conjunction with other sources of information.

Although many researchers indicate the general relevance of their findings to practice, instructions or specific procedures for actually interpreting the effects of important variables (such as class size, required versus elective status of the course, or content area of the course) are not presented. Practitioners who have read Cranton and Smith (1986, p. 127) would be likely to agree that “one must be cautious in extending norms, even to an entire department,” but how many readers have the skills or knowledge needed to make comparisons cautiously among sets of ratings? It is not even known what skills and knowledge are needed for using ratings. Meanwhile,
other sources that debate the merits of such analytical methods as factorial weighting schemes (Marsh, 1987) presuppose the level of skill of the practitioner who would apply them (Abrami and d’Apollonia discuss the importance of dimensionality for interpreting and using ratings in Chapter Seven of this volume).

By combining what we learned from the literature, what we know from our training and experience, and the advice of expert consultants, we were able to construct a large-scale rating system and to demonstrate the validity and reliability of the data we collected. We field-tested each process and procedure, increasing the complexity and scope of the services we offered over time and continuing to the present day.

By far the most difficult problem we encountered (aside from the stresses associated with garnering needed resources and the politics of providing data used in performance appraisal) was deciding how to report the results of ratings to faculty, administrators, and students. In the beginning, we approached the task as one of message design, deciding that printed reports should be readable, with lots of white space and clearly organized information. The available handbooks on ratings, particularly Doyle's (1975, 1983), gave useful suggestions about particular statistics we could use to summarize data. Accordingly, we tried to provide our clients with reports designed for legibility and appropriate analysis of their data.

Nevertheless, conversations with faculty and administrators during the years of development and implementation led increasingly to concern about what users were doing with the information we were providing. We saw that some departmental administrators, who routinely used ratings to make decisions about personnel, evaluation policy, and resource allocation, were not familiar enough with important ratings issues to make well-informed decisions. We regularly heard of personnel decisions that were made on the basis of a single course's ratings and of cases in which workload or difficulty were the deciding factors in such decisions, or in which mean scores, separated by tenths or even hundredths of a point, were accepted as valid indicators of individual differences in teaching performance. Each case is an example of poor ratings practice.

We received many requests from faculty for assistance in interpreting reports, and we discovered that our clients would not or could not use many of the instructions for interpretation that we had provided. Clearly stated disclaimers regarding the limitations of ratings data in particular circumstances appeared to have little effect on the inclination of some clients to use invalid or inadequate data. We found that the lack of resources for teaching improvement often made ratings data practically irrelevant to faculty development.

Designing a better reporting system became increasingly important. Using the systematic development paradigm, we began to look at the skill and informational needs of our clients. The difficulties we experienced looking for sources to inform the design and development of the reporting function led us to survey faculty,
administrators, and teaching-improvement practitioners in five institutions, in order to probe their attitudes toward ratings and their familiarity with fundamental concepts of ratings practice (Fanklin and Theall, 1989).

Our research findings, as well as anecdotal reports from many of our colleagues, suggest that many of those who routinely use ratings are liable to be seriously uninformed about critical issues. For example, among faculty respondents who reported using ratings for personnel decisions involving other faculty, nearly half were unable to identify likely sources of bias in ratings results, recognize standards for proper samples, or interpret commonly used descriptive statistics. Harris's (1990) study of evaluation instruments reinforces our concern. This does not surprise us, but neither does it condemn ratings users. Even though knowledge of ratings may be useful, other issues are the focus of their attention.

A great deal of scholarly attention has been paid to the validity and reliability of ratings as a measure of instructional quality (Cohen, 1981). Considerably less has been given to actual practice, other than the frequency of ratings use in postsecondary institutions. Utilization of ratings is one of the least often studied or discussed issues in the realm of ratings phenomena. There are very few reported observations of ratings users in action in personnel decision making or of the ways in which teaching-improvement consultants use ratings in interactions with their faculty clients. The latter have mainly investigated whether improvement is likely to occur when faculty are helped to interpret their ratings by skilled consultants (Aleamoni, 1978; Levinson-Rose and Menges, 1981; Cohen, 1980). Brinko (1990) takes a closer look at the way clients and consultants have interacted but does not investigate how ratings as such have been used. Similarly, there are few studies of ratings use by students or by those in institutional research.

To illustrate the need to consider the qualifications of ratings users, we offer the case of what to do with less than perfect data. Doyle (1989) rightly notes that academics are liable to "paralysis of analysis," overanalyzing data partly out a need to be safe. His statement that evaluators must work with data that may be less than perfect but that are still good enough reflects the general attitude of many of the most expert respondents in our survey groups. They share the recognition that ratings data are intrinsically both needed and imperfect. It is probably a safe assumption, however, that users who understand how and why data are imperfect are likely to be able to use them safely. Even given the inherently less than perfect nature of ratings data and the analytical inclinations of academics, the problem of unskilled users, making decisions based on invalid interpretations of ambiguous or frankly bad data, deserves attention. According to Thompson (1988, p. 217), "Bayes' Theorem shows
that anything close to an accurate interpretation of the results of imperfect predictors is very elusive at the intuitive level. Indeed, empirical studies have shown that persons unfamiliar with conditional probability are quite poor at doing so [that is, interpreting ratings results] unless the situation is quite simple.” It seems likely that the combination of less than perfect data with less than perfect users could quickly yield completely unacceptable practices, unless safeguards were in place to ensure that users knew how to recognize problems of validity and reliability, understood the inherent limitations of ratings data, and knew valid procedures for using ratings data in the contexts of summative and formative evaluation.

Whether the practices of those who operate ratings systems or use ratings can stand close in inspection has become open to question. It is hard to ignore the mounting anecdotal evidence of abuse. Our findings (Franklin and Theall, 1989), and the evidence that ratings use is on the increase, taken together, suggests that ratings malpractice, causing harm to individual careers and undermining institutional goals, deserves our attention.

**Designing Systems for Use**

Our fundamental assumption concerning the essential function of rating systems is that they should serve the goals of their users. Broadly stated, these goals involve maintaining or improving the quality of instruction provided by institutions or individuals. Accordingly, these goals are best served when rating systems provide useful information to those who make plans, decisions, and choices or take actions directed toward such goals. By examining the goals of ratings users and identifying the events that determine whether the goals can be achieved, we can learn what participants in rating systems (operators and users) must be able to do to achieve those goals. We can also avoid the problems of setting unwanted consequences in motion.

Since the product of a rating system is information, the quality of the product depends on how it is communicated to decision makers and on whether it can be used to achieve their goals. We hope to make our case here by stating the obvious: ratings data are intended to be used, and so ratings results ought to be communicated to users in ways that best facilitate this use.

**Communicating Ratings for Use**

Although people speak of "using" ratings, data begin as responses, on optically scored answer sheets, to multiple-choice statements on questionnaires. Quantitative summaries of raw data must be computed and then transformed into meaningful
depictions of some aspect of instruction, course design, or individual teaching performance. How data are summarized influences both the scope and the content of interpretations. Users rely on interpretations of ratings as evidence when they form judgments of merit or worth. Interpretations of ratings data inform decisions, choices, or plans for action in matters ranging from personnel decision making to faculty development to teaching-improvement activities or such consumer-related issues as course selection.

Using ratings to obtain feedback from students on specific instructional behaviors, in order to inform efforts to improve teaching skills, is an example of ratings use in action. Using ratings to plan institutional strategies for allocating faculty development services is another. Even making negative personnel decisions, which include ratings as evidence of poor teaching performance, is an example (albeit Darwinian) of a strategy to improve instruction by removing all but the most highly rated teachers. The student who selects a course from a catalogue that shows course ratings can be seen as attempting to improve his or her own instructional experience in some way.

To this end, sooner or later in every rating system, the user receives (or constructs) some kind of report on students' responses. With the report in hand, the user forms a working hypothesis about what the ratings mean in terms of the user's own goals. This set of events is perhaps the most crucial in the ratings process, because here is where ratings do their work: at the point where value judgments are made, where attitudes are shaped, and where decisions are made. At the same time, this part of the process is most affected by the quality of the system as a whole because the validity of any judgment rests on the quality of the information the system provides.

The validity of inferences or interpretations should concern those who design and operate rating systems as much as validity and reliability of the instruments used to obtain the data. In a sense, the user is the last defense against bad practice, since the user makes the final decision about whether and how to use any data provided by the system. How use occurs ought to be a very important issue, one for which those who develop rating systems ought to be held accountable. Rating systems must empower users to recognize false or irrelevant information and avoid misconstruing valid and reliable information.

System Components

Taking a larger view of rating systems, it is useful for us to understand utilization in the context of the larger system and its more general events. Information about many of the processes associated with rating systems can be found in (or deduced from) the sources cited in nearly every chapter of this volume (see Chapter Three for a detailed guide).
We describe ratings processes in the context of three activities: establishing systems, operating systems, and using system products. (For the moment, we shall defer discussion of the first.)

Operating a rating system involves the following components:

- Obtaining data (selecting an instrument, where choices among validated instrument are required for services; administering it; performing whatever data-processing and data-management tasks are required for analysis)
- Analyzing data (selecting computational procedures for summarizing, quantifying, or enumerating; producing reports or responses — usually, but not necessarily, in a printed tabular report)
- Providing client-users with knowledge of results (face-to-face consultation printed reports; telephone conversations; any other way of communicating results)
- Other tasks (those associated with data-processing systems and managing organizations; performing error checks and other quality-control measures, systemically and frequently enough to ensure accuracy in collecting and reporting data, budgeting, planning, and training).

How does a system get started, especially if it is complex? Even if one has a list of the processes required to operate a rating system, the list contains no indication of the things one must know or do to perform any of the tasks successfully. Without clear instructions, performing these tasks correctly would be difficult for anyone who was unfamiliar with the sources from which the list was derived. Moreover, without such information, it would be equally difficult for an observer to judge whether a task had been done successfully or at all. In small systems, the same person may be responsible for all these processes; in large, institution-wide systems, people who are never in contact with one another for any other purpose may perform parts of the overall system's work, with little idea of who is doing the rest or even of what the rest is.

We have deferred our discussion of establishing systems to demonstrate that rating systems are more than questionnaires or software. They may be complex entities that need careful planning and management. In other words, we want to demonstrate that the design of a rating system ultimately affects how successfully the system achieves its goals.

Here is a thorny issue, however. In the real world, rating systems are often not systematically designed and implemented. Certainly, an instructor-made questionnaire circulated strictly for obtaining informal feedback ("How am I doing?") requires less systematic consideration than does provision of the organizational resources needed to circulate a mandatory departmental or institutional rating form. Fundamental safeguards and standards for practice are easy to overlook when the system designer assumes that questionnaires and rating systems are synonymous. In the meantime, questionnaires may be disseminated, and data may be collected and analyzed, with little consideration of who is doing what and what the consequences may be. If a mandate to provide rating services on any routine basis is established, the opportunity for careful system design exists. If the mandate itself is embedded in a systematically conceptualized plan for the evaluation of instruction, the system designer's task will be considerably helped. System designers must account for each developmental step in terms of how it affects the ultimate use of ratings data in various domains (personnel decision making, teaching improvement, course
selection, research). Establishing a well-deigned ratings system includes these components:

- Designing the system (conducting research, as necessary, including needs analysis for institution or clients; determining what kind of product or service is wanted and what type of system is needed; performing strategic planning: "drawing board" stage)
- Developing the system (developing and testing prototypes of system components, including questionnaires and data-collection procedures: "working models" stage)
- Implementing the system (putting the system in place; obtaining materials and equipment; training personnel; finalizing system schedules and procedures; building any necessary intraorganizational links, such as making arrangements to use the services of a centralized data-processing agency: "start-up" stage).

Once a ratings system has been established, these activities follow and continue:

- Managing or operating the system (conducting the routine business of administering questionnaires; collecting and analyzing data; reporting results to clients; performing monitoring tasks: "operations" stage)
- Evaluating the system (devising procedures for monitoring accuracy, performance meeting deadlines, utility to clients, and value, in terms of costs and benefits or other indices, in order to improve the design, development, implementation, or management components of the system: "evaluation-revision" cycle).

These processes of establishing a system and operating it are reiterative. Once a system is in place, the activities of improving existing services or creating new ones invoke new cycles of design, development, implementation, management, and evaluation. Additional functions (such as managing public relations, organizational development, and intraorganizational liaisons) lie in a gray zone between establishing and operating a system. We mention these mainly because they are often overlooked but are nevertheless important to the longevity of a rating system. The operating dynamics of the institution greatly affect these functions (see Birnbaum, 1988, for an important discussion of the ways in which colleges work).

**Utilization: A Closer View of Tasks**

Having identified the processes of establishing and operating ratings systems, we now have utilization to consider. Although many of the steps we have described ultimately affect utilization in some way, utilization begins with the user. Here are some of its essential processes:

- Receiving knowledge of ratings data (printed or verbal reports of analyses of data or even unprocessed, raw data)
- Evaluating, quality of the report (establishing its accuracy, particularly course-listing information; determining the validity and reliability of the ratings data; establishing the adequacy of the report for the decision making purpose at hand)
• Interpreting the report (constructing working hypotheses concerning the practical meaning of statistical summaries, in terms of the constructs about teaching or instruction that they are intended to represent; confirming the validity of any working hypotheses provided by the reporting system)

• Synthesizing interpretations of quantitative data with other sources of information to make judgments (making value judgment based on construing interpretations as evidence confirming or disconfirming some specified state of affairs; weighing the ratings evidence against explicit or tacit criteria for a valued state of affairs)

• Taking action based on results (applying judgements of merit or worth to decision making, planning, choosing, revising, recommending, advising, rewarding).

We noted earlier that ratings do their work when someone uses information for a purpose. Doyle (1979, p. 146) says that evaluation researchers “worry about four qualities of any data, including students evaluations: reliability, validity, generalizability, and skulduggery.” To do this we would add that practitioners who design or operate rating systems also must worry about comprehensibility and utility of the information they provide to clients. Recalling the literature, and looking over the processes described in the foregoing lists, we can deduce some of the conditions that are necessary before users will have information tailored to their own purposes.

First, ratings must be collected with validated instruments appropriate to the purpose of decision making. For example, in personnel decision making or course selection, the questionnaire may consist of from three to ten summary items, such as “this instructor compared with others,” “overall amount learned,” or “overall course quality.” Composite scores, computed from factor-analyzed subscales, may also be used, but this is a more complicated process that requires expertise in measurement and evaluation. For teaching improvement, the questionnaire will have twenty or more diagnostic, behaviorally descriptive items, such as “presented information at a rate I could follow” or “created an atmosphere of respect and trust,” as well as any student demographic items that may explain results or contribute to planning for improvement, such as “I believe my background adequately prepared me for this course.”

Second, data must be collected through procedures appropriate to the context in which results will be examined. For example, in personnel decision making, the instructor receives no knowledge of results until after grades are filed. In teaching improvement, the instructor personally administers the questionnaire and tabulates the data.

Third, analyses of data appropriate to the evaluation context must be performed.

For example, standardized scores can be used to show comparisons between courses when personnel decision making is a goal. Course-ratings catalogues should probably be restricted to descriptive statistics that demonstrate how many students said what.

Fourth, results must be presented to users in a format that is suitable for the purpose of decision making and that takes account of the skills and knowledge necessary for users to draw valid inferences from ratings data. For example, in personnel decision making, expert users require only appropriate tabulated, quantitatively summarized data, while novices are likely to need expanded narratives that explain the numbers and their relevance to important constructs in teaching and learning.
Fifth, users must be informed about appropriate procedures for applying interpretations of results to various decision-making processes. For example, in personnel decision making, the user knows that multiple sources of information about teaching performance are required. In teaching improvement, a single set of midterm ratings may be useful and valid.

Note that, despite the increasing detail, the activities still have not been sufficiently described to offer a fully operational guide to practice. It is not clear which things users must do and which must be done by the system. Appropriateness, validity, and effectiveness have not yet been operationally defined. The amount of information increases dramatically as each task is broken down into its components. We now take a closer look at the task of interpreting ratings, and we examine alternative strategies to ensure that users have valid interpretations when they act on ratings.

**Operationally Describing Tasks**

As system designers, our greatest concern, once we have established that we can provide valid and reliable data, should be to facilitate good practice in the use of ratings data in decision making. Selecting or constructing questionnaires, administering them, analyzing and interpreting results, and using interpretations to make decisions are processes that affect the quality of practice to an important extent. These processes can be described in terms of tasks. How these tasks are performed determines the quality of the system's product: the information it produces for decision-making users. Questions of who should have done what versus what actually happened can signal the quality of evaluation practice.

Once we understand the functional relationships among these events, and the conditions that predict success or failure, we can describe what must happen, in what circumstances, and by what standards it can be judged. Thus, we can identify performance problems in the system's components, including problems encountered
by those who operate the system or use its products. When we can identify the skills and knowledge needed by practitioners and users, we can design the system to meet those needs. Once these processes have been identified and are understood, they can be described in terms of tasks, establishing who will do what, when, and how (that is, in what circumstances and to what criterion of precision).

The value of stating a ratings task as a behavior can be demonstrated in these two contrasting statements: "The instructor should administer the questionnaire correctly" versus "The instructor will (1) select a time for conducting evaluations that is during the last two weeks, is before finals, and is not during the same class session as any final examination; (2) designate a student monitor to distribute and collect materials, and instruct the monitor to strictly adhere to the instructions; (3) at the scheduled time, announce that the evaluation will be conducted, and state that the evaluation is important and deserves careful completion; (4) give the materials to the monitor, and leave the room until all evaluation materials have been collected by the monitor; (5) receive neither the completed evaluation packets nor any report of the results until after final grades have been submitted." The first statement is not very useful as a guide to practice. The second statement describes the task in operational terms that define the conditions most likely to elicit valid, reliable data. Constructing such a description not only helps us plan operations but also helps us clarify what we must communicate to our clients to ensure good results.

Many readers may recognize here what has become a routine part of designing or evaluating instructional systems: constructing behavioral objectives. This approach is limited in that it does not address larger issues, such as the value of the goals and objectives themselves, but it does provide a systematic approach to a complex set of phenomena, and it can be particularly useful when what will be evaluated has been designed with explicit goals and objectives in mind.

At this point, having placed the problem of communicating results in the context of the more general processes, and having taken a closer look at the problem of describing the processes in operational terms, as tasks, we begin to plan for designing the utilization process.

**Designing Systems for Valid Interpretation of Results**

The mechanics and style of interpreting ratings appear to vary dramatically across the domains of ratings use, particularly with respect to the role of quantitative information. It is our impression that many teaching consultants employ subjective, experientially based methods of dealing with information, while administrative decision makers may strive to construct empirically based (or empirical-looking) formulas to obtain clear-cut information or to reduce the appearance of subjectivity.
Understanding how users operate should, ideally, influence how a rating system serves them, but considerable research will need to be conducted in this area before any useful generalizations will be available to designers who want to work with users' purposes in mind.

Still, there are some fundamental concepts for using numbers in decision making. To the degree that these concepts are ignored, interpretations of data become, at best, projective tests reflecting what the user already knows, believes, or perceives in the data. Treating tables of numbers like inkblots ("ratings by Rorschach") will cause decisions to be subjective and liable to error or even litigation.

Ratings data are similar in many respects to other survey data, and so ratings do not necessarily require especially esoteric or exotic statistical methods for analysis. Ratings are particularly subject to sampling problems, however, such as not having enough courses on which to base a comparison between two instructors and not involving enough students in rating each course section. Moreover, the fact that classes with fewer than thirty students are statistically small samples means that special statistical methods are required for some purposes. These sampling problems do not negate the value or reliability of student ratings, but we must never overlook such problems when we are using student ratings. Substantially different models for analysis are also required for various uses of the data. Given such problems, there are many opportunities for error in dealing with numbers. Three types of errors come to mind immediately.

The first involves interpretation of severely flawed data, with no recognition of the limitations imposed by problems in data collection, sampling, or analysis. This error can be compared to a Type I error in research-wrongly rejecting the null hypothesis-because it involves incorrectly interpreting the data and coming to an unwarranted conclusion. In this case, misinterpretation of statistics could lead to a decision favoring one instructor over another, when in fact the two instructors are not significantly different.

The second type of error occurs when, given adequate data, there is a failure to distinguish significant differences from insignificant differences. This error can be compared to a Type II error-failure to reject the null hypothesis-because the user does not realize that there is enough evidence to warrant a decision. In this case, failure to use data from available reports (assuming the reports to be complete, valid, reliable, and appropriate) may be prejudicial to an instructor whose performance had been outstanding but who, as a result of the error, is not appropriately rewarded or, worse, is penalized.

The third type of error occurs when, given significant difference, there is failure to account for or correctly identify the sources of differences. This error combines the other two types and is caused by misunderstanding of the influences of relevant and irrelevant variables. In this case, a personal predisposition about teaching style, or the feeling that ratings are popularity contests, may lead a user to attribute negative meanings to good ratings, or to misinterpret the results of an item as negative evidence when the item is actually irrelevant and there is no quantitative justification for such a decision.

How can we conceptualize the problem of ensuring that users do not make decisions or take actions that are based on invalid interpretations of data? In the following example, invalid interpretations are seen to result from either invalid or unreliable data or from lack of skill, knowledge, or necessary information on the part
of the user. The strategy is to make sure that users either have or have access to sufficient skills and information to form valid hypotheses. Valid, reliable hypotheses are those interpretations of ratings data that knowledge, skilled users, with adequate information concerning the present data, would be likely either to produce or concur with.

Let us say that we state our goal in the following way: "The user will make decisions that are based only on valid, reliable hypotheses about the meaning of data." In this case, the user should receive or construct working hypotheses that do the following things:

- Take into account problems in measurement, sampling, or data collection and include any appropriate warnings or disclaimers regarding the suitability of the data for interpretation and use.
- Do not attempt to account for differences between any results when they are statistically not significant (probably < .05).
- Disregard any significant differences that are merely artifacts of measurement (for example, small differences observed in huge samples, which can technically be significant but are unimportant).
- Account for any practically important, significant differences between results in terms of know, likely sources of systematic bias in ratings or reliably observed correlations, as well as in terms of relevant praxiological constructs about teaching or instruction.

The user should also refrain from constructing or acting on hypotheses that do not meet these conditions.

Since the operator provides the user with summaries of results, the role of the operator is crucial in ensuring that valid interpretations are possible and, better yet, likely. Some of the necessary conditions for operators can be inferred from the preceding list. Operators should ensure that all ratings data are sufficiently tested for validity and reliability before they are communicated to users; provide warnings to users concerning possible errors or problems in measurement; provide users with summaries of results only when the data meet the validity and reliability criteria appropriate to the decision-making purpose of the users; and provide users with valid interpretations or any necessary direct assistance for constructing them.

From this perspective, the system designer's task goes beyond testing the validity and reliability of all ratings instruments to be used in the system. The designer must also consider the tasks operators and users will perform and must do whatever is necessary to make users likely to refrain from making decisions based on invalid, unreliable data. The designer must be sure that procedures for testing and reporting the validity and reliability of data for each rated course and section are implemented. Furthermore, the designer must ensure that the user is either explicitly instructed or advised on procedures for determining validity and reliability and is informed of the necessity for using those procedures. For example, the designer can
take measures to see that the user recognizes and rejects invalid or unreliable data, knows the criteria for validity and reliability, recognizes sources of sampling errors and other measurement errors, and accounts for known sources of error in any interpretation. This means that the designer must establish procedures to ensure that users are trained in, given instructions for, or provided with consultation on constructing hypotheses or else are provided with valid interpretations of ratings results and instructions for their use. The designer must also evaluate the quality of the decisions and actions of operators and users with respect to the quality of data and of the interpretations and revise the system as necessary.

Update on a Development Project.

We do not mean that there are rigorous, specific, known methods for designing rating systems, but only that there are useful heuristics for dealing with analyzing processes and tasks in complex systems, heuristics that can be productively applied to the task of communicating ratings data or to any other aspect of rating-system design. We continue to find that the systematic approach we have describes here is an essential tool for improving the services we offer.

We currently provide clients with specific report forms designed for use in teaching improvement, course selection, and personnel decision making. In our current ongoing system-development process, we have devised and disseminated report forms that meet the guidelines suggested in the literature. We have found, however, that many difficulties attend their use. As a result, we have entered a new development phase in which our objective is to learn more about the needs and skills of our decision-making users, in order to produce reports that will make their uses of ratings data as effective, efficient, and fair as possible.

During this stage, we plan to radically redesign the report that we provide to personnel decision makers by experimentally offering several alternative treatments for summarizing and explaining data. Each treatment will meet the general conditions established in the literature but will vary in format (graphic, numerical, narrative) and scope (embedded, individualized, interpretive comments versus general instructions). We will use interviews and surveys to probe the reactions of our users to these reports and will attempt to understand what characteristics of reports facilitate good practice.

Conclusion

We are convinced that those who develop and operate rating systems have an obligation to make the systems support good practice. It is not enough to provide valid, reliable information if the users are not prepared to make use of it. The rating process and rating systems are anything but simple. Individuals and institutions should be concerned about more than the wording of questionnaire items or how many points ratings will be worth in decisions on promotion and tenure.

We can assume that many of those who come to important ratings tasks will be relatively unfamiliar with at least some of the critical issues involved and will not
have the time or even perceive the need to become qualified practitioners and users. The process we have employed to describe rating systems may seem tedious to some readers or merely commonsensical, despite its trappings of systems rhetoric. Our experience has convinced us, however, that the obvious is likely to be overlooked in the absence of tools for systematic development.

What can we do to improve practice, in these circumstances? Our feeling is that seeing the system in terms of its goals and what is needed to achieve them gives us an opportunity to discover a variety of alternative strategies for helping system developers, operators, and users do their various tasks. Allowing only skilled operators and users to participate in the system, or providing training, would certainly help establish good practice.

Because nearly everyone who is a client of rating systems is involved in other full-time work in a wide variety of academic disciplines, finding or training skilled users is often impractical. In fact, it is axiomatic that training is not always the only or even the best solution to every performance problem. Alternatives to training include providing detailed, step-by-step instructions ("job aids") and providing supervision and consultation from qualified peers or expert practitioners.

Rating systems themselves need periodic evaluation. Determining how effectively ratings are doing the jobs for which they are intended, and what other unplanned consequences attend their use, is essential to the integrity of any evaluation process but particularly to those that include consideration of ratings for personnel decision making. We need to understand current practice in operational terms. We cannot rely on having valid, reliable items, appropriate analysis, or useful reports if practitioners and users are unable to fully execute the tasks that underlie validity, reliability, appropriateness, or utility. It would be worse still if these persons were unaware of these requirements or did not accept their importance.

We regret having only scratched the surface of the problems of ratings use in our rush to make an important point: that the proper use of ratings should never be taken for granted, without strict consideration of the quality of the ratings data or the qualifications of users. For example, we barely mentioned tasks associated with policy and regulation, yet they are crucial to success. We totally omitted consideration of qualitative approaches to evaluating teaching and how they relate to ratings use. We even ignored one of our own fundamental assumptions: that without the opportunity for improvement of teaching skills, ratings are punitive.

Because the scope and complexity of rating systems will vary from site to site, in terms of who does what, how, and to what end, we are unlikely to find a single solution to improving practice. When we understand ratings systems as objectively described tasks, however, we can talk about what qualifications, training, or consultation people must have in order to perform the tasks effectively. Ultimately, we can also use this view to determine how well our rating systems are working, whether they are achieving their goals, and what we can do to improve them.
References


Jennifer Franklin is senior research associate at the Center for Applied Social Research and manager of the Office of Instructional Evaluation at Northeastern University.

Michael Theall is associate professor and director of the Center for Teaching and Learning at the School of Education, University of Alabama, Birmingham.
Appendix G

Theall & Franklin (1991)
In a student rating report, a mean may have little meaning in the absence of reliable, valid data and appropriate bases for comparing it to some standard. To use ratings result effectively and accurately, guidelines for understanding data are crucial. Why, then, are they so often absent?

Using Student Ratings for Teaching Improvement

Michael Theall, Jennifer Franklin

As many contributors to this volume point out, the most effective kind of feedback for teaching improvement combines valid and concrete information with the assistance of a knowledgeable person in a nonthreatening environment. Also acknowledged by writers and practitioners in evaluation and teaching improvement is the fact that student ratings of instruction are the most available and most commonly used source of feedback for both reaching improvement and personnel decision making (Seldin, 1989). Unfortunately, student ratings are also the kind of data most frequently misinterpreted and misused by faculty and administrators (Franklin and Theall, 1989). However, this finding is tempered by the fact that the same research also found that instructional specialists and teaching improvement practitioners (some
of whom were also faculty) had significantly higher levels of knowledge about evaluation practice and the use and interpretation of student ratings than did faculty colleagues who were not involved in teaching improvement activities. Nonetheless, the tasks of interpreting and using ratings are not simple and are often confounded by the various levels of complexity imposed by the contexts in which they are used (Theall and Franklin, 1990).

In Chapter Four of this volume, Peter Gray offers an approach to gathering and using data for teaching improvement. In Chapter Five, Karron Lewis considers various sources of data, including student input from small group instructional diagnosis. The purpose of this chapter is to provide guidelines for using student ratings for teaching improvement, with special attention given to the teaching consultation process.

We begin by grounding the process of reaching improvement consultation in the theoretical context of dimensions and specific behaviors associated with effective teaching. We do this to give the practitioner and faculty client common ground for evaluating teaching performance. When consultation is understood as a learning process (for both client and consultant), this framework makes instructional sense as well. We concentrate primarily on the following issues: the dimensions and behaviors of teaching that can be evaluated using ratings, useful questions to include in the evaluation effort, technical characteristics of valid and reliable questionnaires and items, use of student comments, useful analyses and reports, interpretation of results, and “other issues” for the consultant who uses ratings feedback.

The Dimensions and Behaviors of Teaching

There has been considerable investigation of the dimensionality of teaching. Conclusions have been generally similar, identifying teaching (and its evaluation) as multidimensional (Cohen, 1981; Feldman, 1976; Marsh, 1987) and specifying the dimensions as similar. Recently, Abrami and d’Apollonia (1990) identified twenty-four dimensions: (1) stimulation of interest, (2) enthusiasm, (3) knowledge of the subject, (4) intellectual expansiveness, (5) preparation and organization, (6) clarity and understandableness, (7) elocutionary skills, (8) class level and progress, (9) clarity of course objectives, (10) relevance and value of course material, (11) relevance and usefulness of supplementary materials, (12) workload, (13) perceived outcome, (14) fairness of evaluation, (15) classroom management, (16) personality characteristics, (17) feedback, (18) encouragement of discussion and diversity of opinion, (19) intellectual challenge and encouragement of independent thought, (20) concern and respect for students, (21) availability and helpfulness, (2) overall course, (23) overall instructor, and (24) miscellaneous items.

Many of these dimensions are not appropriate for use in personnel decisions, but each dimension is potentially important for teaching improvement. While some dimensions are general (for example, 22 and 23), others are more specific and can be investigated in detail. An overall rating of teaching performance (23), although summary in nature, is not necessarily the “average” of more specific ratings. In fact, ratings on specific teaching skill areas may be very different from overall ratings. The
teaching consultant can help faculty to identify areas or dimensions in need of attention and can also show how items that probe specific dimensions of teaching can relate to overall ratings.

An effective way to use student ratings to assess performance in each of the dimensions listed above is, as Murray (1983) has suggested, to approach teaching in terms of "low-inference" behaviors, that is, specific unambiguous behaviors that can be isolated and observed without interference from the value system of the observer or other confounding factors. The benefit of this approach to teaching improvement is that when important teaching skills are assessed through low-inference behaviors, areas of weakness become identifiable and amenable to change. Low-inference behaviors can often be learned without difficulty. Thus, faculty can quickly gain skill in these behaviors.

As suggested elsewhere in this volume, the initial targets for teaching improvement should be those behaviors that can be clearly identified and addressed. Less likely for success are attempts to alter long-standing patterns of working or cognitive style, strongly held philosophical beliefs, orientations to one's field or toward teaching, and personality characteristics. In many instances, as George Geis suggests in Chapter One, it would be presumptuous to suggest changes in these areas based on the results of one evaluation.

Useful Items to Include in Ratings Questionnaires

If the goal is to determine whether behaviors associated with effective teaching occur and to measure how often, a ratings questionnaire must address those behaviors. Items that probe behaviors generally associated with effective teaching as well as items specific to content area instructional techniques (such as labs or case studies) give useful information when students agree in their responses.

However, students may disagree in their perceptions of what happens in a classroom. What is helpful for one student may not be effective for others. Proper use of ratings feedback in consultation requires an understanding of the sources of disagreement as well as the common ground. Instruction is not a unilateral process. Students should be active participants in the teaching-learning process. Many characteristics of students, such as their backgrounds, prior preparation, general attitudes toward the content area of the course, achievement level (grade point average [GPA]), and expected course grades, can be helpful in understanding responses to items that probe low-inference teaching behaviors. For example, students may disagree strongly in their responses to an item that probes whether the pace of instruction is appropriate for their skill level. If the consultant can determine that
students with low GPAs tended to be dissatisfied with the pace compared with those with middle to high GPAs, then the consultant has an important clue about what is happening in the course and how instruction can be improved for the low-GPA students.

It is also helpful to use two or three very general items that probe satisfaction with the course, teacher and amount learned (Cashin and Downey, 1991). These items can help the consultant and faculty member form a more reliable general impression of students' feelings than can be achieved by average (unweighted results from more specific, individual items.

Technical Considerations

Teaching improvement consultation requires ratings instruments that contain a combination of general and specific items probing aspects of instruction associated with effective teaching. The securing of such questionnaires and items is more problematic.

There are three ways to address the instrumentation problem. The first I to employ a questionnaire that has already been validated and extensively used (for example, the IDEA form developed at Kansas State University, the Teaching Analysis by Students form developed at the University of Massachusetts, the Teacher-Course Evaluation Project Questionnaire developed at Northeastern University, the Course/Instructor Evaluation Questionnaire from the University of Arizona, and the Student Evaluation of Educational Quality form developed at the
University of California). The second is to develop and validate a questionnaire locally, although this is a difficult and time-consuming process (see Arreola and Aleamoni, 1990; Berkj, 1979; Theall and Franklin, 1991). The third strategy is to tailor-make a questionnaire for selecting items from an item bank. In each case, the instrument constructed should contain items that allow clear interpretation, probe only one issue at a time, and state the issue in unambiguous terms. As suggested above, low-inference items are useful for examining teaching skills. For example, in dealing with the dimension of clarity, Murray (1983) suggests questionnaire items that investigate the following teaching activities: (1) providing frequent examples, (2) defining new or unfamiliar items, (3) repeating difficult ideas, (4) using graphs or diagrams, (5) pointing out practical applications, (6) suggesting ways to memorize, (7) writing key terms on the board or overhead screen, and (8) answering questions thoroughly.

The strategy of local development of questionnaires should remain a major area for concern. For while the temptation to locally develop such forms is great, the end result is often invalid and unreliable data, especially when the faculty develop the forms without expert assistance. In a study of the evaluation items chosen by faculty, Ory and Wieties (1991) reported that the ten items most frequently chosen from a cafeteria-style item bank included questions about instructor preparedness (ranked first), instructor knowledge (third), grading practices (fourth), and enjoyment of teaching (eighth). From a utilitarian viewpoint, these are items of dubious value, and in fact, items on preparedness and knowledge are generally considered invalid because students, at best, can only report their own impressions of an instructor's
apparent subjects expertise. Presumably, a subject matter expert would be better qualified to rate preparation and knowledge. Only two of the frequently chosen items in Ory and Wietjes’s study approached low inference: “How well did exam questions reflect content and emphasis of the course?” (fifth) and “The instructor stated clearly what was expected of students” (sixth). Thus, faculty item choices not only appeared to yield less information about specific teaching behaviors that is needed for teaching improvement purposes but at the same time provided other information of questionable value.

Wording of Items and Response Options

The wording of items and response options should be clear and consistent. A good combination includes questions phrased as statements of behavior and either a frequency or an agree-disagree response scale. For example, the question stem might state, "The instructor presented information at a rate I could follow." Note the use of the personal pronoun "I" rather than the less precise word "students." Individuals do not always know the feelings of other students and therefore questions should not ask them to estimate the opinions of others. With either the frequency or the agreement scale, this item can provide a clear picture of the effectiveness of the instructor’s pacing for each student and, by inference, for the group, because most students are likely to understand the question in the same way. But if the item were phrased, "The instructor paced the instruction appropriately" or "The instructor presented information at an appropriate pace," variation in students’ definitions of "appropriate pace" would make interpretation more difficult. Moreover, if the item read, "Rate the appropriateness of the instructor's pacing," and the response scale covered five points from "excellent" through "average" to "poor," two definitions come into play: one for "appropriateness" and a second for the response adjective chosen. Clearly, this item is likely to be even less reliable than the previous example.

The preceding discussion illustrates how item construction can affect the overall usefulness of a questionnaire. The need for reliable information argues for the use of a validated instrument or the consultation of measurement and evaluation experts. If an instrument is locally developed, then careful field tests, analysis, and validation are always required.

Student Comments

Almost all student rating systems allow the inclusion of written comments. Most faculty pay close attention to these comments, and some administrators want to see them for use in promotion and tenure decisions. Narrative comments are given great
weight. Our experience has been that about 10 percent of a class responds with narrative comments unless an extreme situation arises, whether good or bad. In the extreme cases, comments match quantitative results in terms of frequency and intensity, but in more "normal" situations (that is "average" ratings in courses with normal distributions of scores) comments usually come from either the very satisfied or the very dissatisfied. It is easy to overinterpret these comments. It is also dangerous because they represent so small a percentage of the class. If we are unwilling to accept quantitative data from fewer than half the students, (the minimal standard described in Table 6.1), we should not be willing to make decisions based on a 10 percent sample simply because the information is in written rather than numerical form.

This minimal standard of 50 percent does not mean that student comments are unusable for teaching improvement. Indeed, student comments can provide very valuable insights into classroom processes and activities or into teacher behavior. But they should guide further investigation and be used in conjunction with other kinds of data, rather than solely to determine a course of action or a decision. Articulate, insightful student comments can be a potent tool for illustrating what quantitative data show in more abstract terms.

**Analyses and Reports**

At minimum, analysis of evaluation results for teaching improvement should include descriptive information (distributions of responses by item), measures of central tendency (mean, median, standard deviation), and a direct estimate of error such as confidence intervals for means. Although global items, factorially derived, or other composite scores for subsets of items that probe several aspects of a single topic are often recommended for personnel decision-making purposes, they are probably less useful for providing feedback for teaching improvement unless accompanied by more specific information.

While the purpose of teaching improvement evaluation is not to compare the performance of individuals, interpretation of results is clearer if performance can be measured against some standard. Centra (1979) has suggested that comparisons made among faculty can help motivate poor performers to improve. Certainly, comparisons among course offerings over time for a single teacher can demonstrate the effects of efforts to improve. Analyses comparing individual results to some norm usually involve t-values or other standardized scores, or percentile-ranked means with confidence intervals, and they require a sufficiently large data base of course evaluations. Generally, teaching improvement purposes do not require that individuals be so rigorously compared to standards derived from some group or even from their own past performances.

For comparisons among teachers, the most logical standards are the norms for the group (course, department, college, institution, or national data base). But recent research has emphasized the importance of avoiding comparative judgments when the bases for comparison are ill-defined. Cashin (1990) has shown consistent differences in ratings as a function of disciplinary differences, and Franklin and Theall (1991) report large and significant negative effects of class size on both student ratings and
student performance. While these findings argue, once again, for caution in the
interpretation of results, they do not mean that interpretation is impossible. Good
reports, useful guides, and knowledgeable assistance are required for this purpose.

<table>
<thead>
<tr>
<th>Class Size (N)</th>
<th>Minimum Acceptable Response Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 - 20</td>
<td>80 percent</td>
</tr>
<tr>
<td>21 - 30</td>
<td>75 percent</td>
</tr>
<tr>
<td>30 - 50</td>
<td>66 percent&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>50 - 100</td>
<td>50 percent&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>&gt; 100</td>
<td>50 percent</td>
</tr>
</tbody>
</table>

<sup>a</sup> 75 percent recommended

<sup>b</sup> 66 percent recommended

Interpreting Reports

When reports are produced, they should present information clearly and should be
accompanied by guides for interpretation and use. All too often, evaluations are
processed (by default) in a campus computer center and the only feedback received
by faculty is a set of numbers on "green bar" computer paper.

Steps can be taken by individuals and practitioners to help ensure valid
interpretation of data. First, examine the size of the class. The number of students
responding, and the ratio of these figures. Table 6.11 presents a general set of guidelines.

Second, examine each item to determine the percentage of omitted responses. Even in a good sample, some items may have been left blank because they did not apply or because students were unsure of their responses. Reduce the response ratio for the item by the percentage of blank or omitted responses and reconsider its usefulness.

Third, review the distributions of responses and the mean scores and standard deviations of individual items. Some reports of results provide this information for all items regardless of whether the information is meaningful. For example, "workload" and "difficulty" are often misinterpreted because some value is attached to heavy workload or extreme course difficulty under the automatic assumption that "heavy and hard" results in poor ratings or that "light workload" and "easier" levels of difficulty signal poor teaching. These statements have no stronger empirical basis than does the argument that good ratings can be "bought" with high grades or that teachers who are "easy" receive higher ratings. Franklin and Theall (1991) have shown that there is a significant and positive relationship between challenge and ratings. Students gave higher ratings and demonstrated higher levels of achievement in courses that were rated as "more difficult" because they had achieved more. They also gave high ratings to courses and instructors when courses were less difficult as long as they felt that they had learned something. And in the several thousand courses in the sample, there was no evidence whatever to support the "grade inflation" hypothesis. If there were attempts to "buy" higher ratings with higher grades, students were having no part of it.

The mean is the arithmetic average of student responses for an item. Means for ratings items often range from 1 to 5. Check the response options on the evaluation form used to be certain of correct interpretation. One characteristic of
student ratings that has been reliably observed during the last several decades of
teacher-course evaluation is the tendency of students to rate instructors favorably.
Thus, the “average instructor” is typically rated around 3.5. This produces a skew
toward the high end of the scale, and the otherwise bell-shaped curve of instructor
ratings becomes “asymmetric.”

The standard deviation for individual items is an index of agreement or
disagreement among student raters. Perfect agreement yields a standard deviation of
0. Deviations of less than 1.0 indicate relatively good agreement in a 5-point scale.

Deviations of ½ and higher indicate that the mean may not be a good
measure of student agreement. This situation may occur when opinion in a class is
strongly divides between very high and very low ratings or, possible, is evenly
dispersed across the entire response scale, resulting in a mean that does not represent
a “typical” student opinion in any meaningful sense Because students vary in their
needs, just as teachers vary in their skills, a teacher may be “among the best” for
some and at the same time “among the worst” for others. A mean of 3.0 or 3.5 cannot
be constructed to represent “average” performance in the sense of meddle-range
performance when he mean is simply an artifact of strong disagreement among
students. The standard deviation is therefore an important source of information about
student opinion.

While means with small deviations can be understood to represent student
opinion in the semantic sense (for example, “there was good agreement among
students that this course was among the best they had taken”, means alone are not
recommended for comparisons among courses. The majority of means fall within one
standard deviation above or the grand mean (the “mean of means”), and few people score extremely high or low. To make distinctions between scores in the midrange between very high and very low means is to risk overinterpreting chance differences in scores. For personnel decision making, use $t$-scores, percentile-ranked groups, or other appropriated measures of relative performance. In teaching improvement, use means only to locate the individual’s performance in context. Obviously, improvement efforts are not necessary if an individual is performing well with respect to the norms for the department or school, barring the unlikely situation that the unit norms are themselves depressed in such a case, improvement efforts on a larger scale may be warranted.

**When Good Teaching Is “Average”**

As noted earlier, one characteristic of student ratings that has been reliable observed during the last several decades of teacher-course evaluation is the tendency of students to rate instructors favorable. Thus, mean scores that are equivalent to “above average” in the original response scale may be grouped among the lower middle in percentile-ranked groups or may be distant from the mean in a $t$ or other standardized score.

In effect, any score in a sample group in which no instructor was rated less than “better than average” is not a poor or even a mediocre score, even if its percentile group equivalent group is low. Here is a concrete example: All faculty in a department receive overall ratings of 4.0 or higher (5.0 is the best possible rating).
The norms on this item for the school and the institution are 3.7 and 3.5, respectively, so Instructor X (whose ratings is 4.2) is considered, in the language of the item, "better than most instructors I have had." But instructor X falls in the seventeenth percentile of the department, a ranking that carries a very different meaning from what the student raters intended. If the range of scores is very small, percentile rankings ranging from top to bottom represent trivial difference in mean score equivalents. In such circumstances comparisons of ratings is not an appropriate method for making distinctions among instructors. Other sources of data are essential.

**Other Considerations**

Although more than three decades of research have shown student ratings to be quite dependable, there are a few predictable sources of systematic variation that should always be considered in comparing scores. For example, here are a few conclusions from the research (see Arreola and Alemoni, 1990; Cashin, 1990; Centra, 1979; Franklin and Theall, 1991; Marsh, 1987): (1) Students tend to rank instructors teaching small classes (less than ten or fifteen) most highly, followed by those with sixteen to thirty-five students and those with over one hundred. Thus, the least favorably rated are classes with thirty-five to one hundred students. (2) Students tend to give slightly higher ratings to their majors or electives than to courses taken to fulfill a college requirement. (3) Courses in the humanities tend to be rated more highly than those in the physical sciences. (4) Student ratings can be biased by failure to adhere to instructions for administering a questionnaire. For example, failure of the instructor to leave the room during administration, failure to preserve student
anonymity, administering of the evaluation during finals, and use of prejudicial introductory remarks can invalidate ratings results. (5) time of day and scheduling factors appear to have little or no influence on ratings. (However, there may be systematic differences in who attends classes at particular times, which may have some impact on ratings.) (6) Academic ability of students as measure by GPA has shown little relationship to the ratings that they give. That is, “poor” students are just as appreciative of good teaching as “good” students, while good students are just as critical of poor teaching as are less able students. And both types of students can easily tell the difference. (7) The practices of giving students unearned good grades have been demonstrated to have virtually no payoff of raising ratings in the long run, provided that ratings are anonymous.

Other Issues for Consultants

Knowledge of the literature on student ratings and of relevant issues of measurement and statistics is important for teaching consultants because helping faculty to interpret evaluation reports is one of the most common kinds of assistance that consultants provide. But there are other, equally important data-related issues for teaching consultants. The following sections outline some of these issues and explain why these kinds of information are valuable.

Establish Procedures to Document the Consultation Process and Its Results. For teaching improvement, documentation serves both the client and the consultant. For example, both consultant and client have a record for future reference, and the consultant can use the documentation to improve consultation skills.
Help the Institution to Use Ratings Accurately and Effectively. The teaching improvement consultant is often the only person available to provide information about the valid uses of ratings in the personnel decision-making process. Many faculty and administrators believe that where ratings forms are concerned, more is better, that is, more items provide a better basis for personnel decision making. However, ratings used for teaching improvement must first be confidential. The detail required for effective use of ratings in teaching improvement can be easily misinterpreted in the context of personnel decision making. The differences between contexts of application pose a particular problem for the consultant who is asked by a teacher to provide an administrative decision maker with copious of the results obtained with a diagnostic ratings form.

A standard diagnostic questionnaire might, for example, contain items about leading discussions. However, in a very large lecture class of one hundred or more students, it is very difficult to conduct useful in-class discussions regularly, if at all. If the response scale for the questionnaire ranges from “almost always” to “almost never,” students might report that the instructor “rarely” or “almost never” conducted meaningful discussions. This item can be ignored by the instructor, who correctly views the responses as an artifact of the teaching situation. But if the report were made available to others unfamiliar with the class and they did not note the class size, then a serious interpretive error could result to the detriment of the instructor.

Be Aware of the Usefulness of Other Data. Just as specific questions can be useful for gaining an understanding of the behavior of the teacher, other data are important for understanding the students and the situation. Conclusions drawn about teaching in the absence of this information are reached with much more difficulty and are prone to more error, thus, teaching improvement evaluation should always consider these other sources of data. Knowledge about student motivation, prior preparation, class, and GPA, for example, can help the consultant understand why
some students report that they cannot follow lectures or that they find tests extremely difficult.

Another important reason for collecting information about students is that it may prevent misinterpretation of results about faculty. In fact, for the protection of faculty, ratings items about students should also be considered in interpreting ratings results for personnel decisions. Consider the following case taken from actual data.

Professor X requested help from a teaching consultant in interpreting recent student ratings. The consultant noted the following pertinent information about the teaching situation and its evaluation: (1) ratings were somewhat lower than usual on the overall items, (20 ratings were low on items relating to testing, pacing, relevance, and clarification of problems, (3) many items had unusually high standard deviations, and (4) the course's workload was considered "heave" and the course rated "more difficult than average." All in all, ratings were considerably lower than usual for this instructor and were marginal in comparison to the norms for the department. The overall ratings of the course "were to appear in the ratings catalog and were probably going to be considered in an upcoming tenure decision. Professor X was concerned about whether this evaluation would help or hinder a favorable decision. The consultant responded that the effect might depend on whether the results would be fully interpreted.

After reviewing the results, Professor X and the consultant decided to inspect other information. A check of student demographics revealed the following: (1) About 40 percent of the class were seniors, 40 percent were freshmen, and the rest were equally divided among the other classes. (2) These percentages were similar to the distributions of responses on the items about prior preparation of students, difficulty, pacing, and, in fact, most of the specific items with depressed scores. Also, a review of teaching load revealed that since employment four years prior, this teacher had taught only upper-level or graduate courses. This conclusion (borne out by further analysis of the evaluation results) was that he succeeded with upper-level students, but the lower-level students had difficulty keeping up and thus were negative in their opinions.

Are the depressed ratings the teacher's "fault" alone or are they also related to a scheduling coincidence and a curriculum problem? One could make a case for the latter explanation and lessen the effect of this set of ratings on an overall assessment of the teacher's performance. As a result of the evaluation, some departmental changes might be made in the requirement for this course or its placement in the curriculum. Also, the teacher might decide (if possible) to focus on upper-level courses, or to work with a colleague who has been effective with beginning students, or to work with a teaching consultant to investigate the situation in more depth and to develop strategies for teaching lower-level students. Although this consultation began with the client's concern about the effect of ratings on tenure prospects, the information provided by the evaluation also created a starting point for improvements that could benefit the teacher, the students, and the institution.

When evaluations are used for promotion and tenure, the accepted rule is that no single evaluation should be considered adequate for decision making. In the above case, the instructor's overall record of good evaluations lessened the negative impact of one particular set of results. But in teaching improvement, consultants may be tempted to provide guidance based on a single set of results. Practitioners would do well to remember L. Hommedieu and Menges's (1990) caution about the quality of samples used in studies of ratings feedback. They reported that, in general, the conclusions of feedback studies were based on very small, often inadequate, samples and that generalizations arising from classroom studies should be carefully made. Likewise, teaching improvement practitioners would be well advised to avoid making generalizations about teaching performance based on the results of one application of a ratings instrument. Consultations based solely on a single sample of data should focus on understanding what happened in that particular class rather than attempt to characterize the teacher's typical performance. However, if the consultant has
additional data from classroom observations, interviews with students, and other modes of evaluation, a single set of ratings may be useful for illustrative purposes. The important point is that decisions about useful strategies for improvement should not be based on a single sample of ratings data.

Summary

Teaching improvement efforts require valid and reliable information, which students can provide efficiently and effectively through the use of ratings questionnaires. In order for the data to be useful and for the processes of improvement to succeed, a series of careful steps must be followed. In particular, consultants and faculty need to know how different evaluative purposes affect evaluation results and they must be able to interpret and use the data at hand. Ill-informed decisions not only waste time and resources but also have negative effects on the faculty, the students, and the institution.

References


Theall M., and Franklin, J. "A Low-Threat, High-Impact Method of Developing a Student Ratings Questionnaire in a Department or College." Workshop presented at the 12th annual sharing conference of the Southern Regional Faculty and Instructional Development Consortium, Atlanta, Georgia, February 1991.

Michael Theall is associate professor and director of the Teaching and Learning Center at the School of Education, University of Alabama, Birmingham. Jennifer Franklin is associate director for evaluation and senior research associate in the Office of Instructional Research and Evaluation of the Center for Applied Social Research at Northeastern University, Boston, Massachusetts.
Appendix H

Recruitment Letter for Control Group
May 8, 2001

Dr. Paul Douillard
Vice President for Academic Affairs
Caldwell College
Caldwell, NJ 07006

Dear Dr. Douillard:

I am writing with an invitation to participate in a research project on student course evaluations. Few topics in higher education generate as much dialogue as student evaluations of teachers (SET). The study is designed to ask participants to respond to various questions regarding end of semester student course evaluations. Participants will be asked to complete an email survey and send it to me electronically. They will then be asked complete the SET survey again electronically. Can you identify up to four faculty, chairs, deans and promotion committee members who might be interested in, and benefit from participating in this study?

Please forward this information to your colleagues with my email address. They may also call or write me regarding the study.

Sincerely,

Michael A. Rossi, Jr.
Michael A. Rossi, Jr.
Centenary College
400 Jefferson Street
Hackettstown, NJ 07840
908.852.1400 x2117
908.813.1984 (fax)
mrossi@centenarycollege.edu
Appendix I

Franklin & Theall (1989) Survey
Original Survey by Franklin and Theall (1989)

INSTRUCTIONS: Part I.

Using the responses below, write the response that most closely matches your reaction to the statements in items 1 through 41 next to each item in the space provided.

A = strongly agree  
B = tend to agree  
C = tend to disagree  
D = strongly disagree  
E = uncertain, no opinion

1. There is rarely good agreement between the ratings of students and other measures of teaching competence (such as administrative reviews, peer observations, and interviews with students conducted by independent consultants).

2. Student rating questionnaires properly constructed and administered can do a good enough job of measuring teaching effectiveness to warrant their use for personnel decisions.

3. Student ratings can’t really give a teacher the kind of feedback needed for improving the quality of instruction s/he provides.

4. Very specific items dealing with a teacher’s behavior and skills usually correlate more highly with student achievement than do broad, global items such as “this instructor compared to others.”

5. Really good, but tough teachers will tend to receive lower ratings and will only be appreciated by students years later when real life experiences have provided a new perspective.

6. Fifteen students in a given course are more likely to agree with each other in their rating of an instructor than are fifteen of the instructor’s colleagues rating the same course.

7. Even a few written comments from students can often tell more about a course than a full page of statistics from a student rating questionnaire, no matter how many students responded to it.

8. Discussion, lecture and laboratory classes are so different that there is no valid way to use student ratings to compare the instructors who teach them.

9. Student ratings used for personnel purposes should not be returned to the instructor until after final grades have been submitted.

10. When student ratings are used for promotion and tenure decision-making, it really doesn’t matter who administers the course evaluation.
11. Students should be required to sign their answer sheets.
12. The instructor should leave the room during the administration of a student rating questionnaire.
13. When ratings are mandatory, every course/section of every instructor should be evaluated every academic term.
14. Telling students that results will be used for personnel purposes tend to cause students to be more critical and thus lower ratings.
15. There is no predictable relationship between student ratings and student achievement.
16. Students are more likely to give slightly higher ratings to their majors or electives than to courses taken to fulfill a college requirement.
17. The ratings obtained from “good” students should be given more weight than those of students who aren’t doing well, i.e. the higher the grade point average of the rater, the more likely valid that rater’s responses.
18. A student’s age and sex tend to exert predictable and systematic influences on the ratings s/he gives.
19. There are no predictable differences between the ratings given to lower level (beginning) courses and those given to upper level courses.
20. Students who expect higher course grades tend to give more positive ratings than students who expect lower course grades.
21. Charismatic, entertaining teachers tend to get better ratings on “overall” items (e.g. “this instructor compared to others” and “this course compared to others”) than less “flashy,” but otherwise better teachers.
22. An otherwise poor teacher can get higher ratings by lenient grading.
23. Generally, a rating of heavier than average workload and/or difficulty is associated with lower than average ratings of the teacher’s performance based on “overall” item scores.
24. The content areas of courses is associated with predictable differences in the ratings students give.
25. Class size has no predictable influence on student ratings.
26. Administrators should usually give more weight to student rating items dealing with an instructor’s specific teaching skills (such as communication, rapport, testing, and grading) than to broad, global items such as “this instructor’s teaching effectiveness compared to others.”
27. Properly constructed, administered, and analyzed student ratings are usually sufficient as the sole source of information about teaching effectiveness for the purpose of personnel decision-making.
28. Assuming that an adequate number of students respond to have a reasonably representative sample, the smaller the classes, the more classes are needed to determine an instructor’s average teaching performance.
29. The number of courses needed to provide a reasonable basis for understanding a teacher’s performance depends largely on how the rating results are to be used.
30. Given a valid questionnaire properly administered, the ratings obtained in a single large lecture course (e.g. 100 students) would be an adequate sample of
student opinion for use in personnel decision-making, provided that at least 75% of the students responded.

31. The proportion of a class that rates an instructor is not as important as the total number of raters.

32. The best measure of overall teaching performance is obtained by averaging the scores for each item on the questionnaire into one ‘grand mean.’

33. When interpreting ratings results, one way to help control the influence of bias in student ratings due to factors beyond the instructor’s control is to consider scores for any available, relevant comparison groups.

34. For personnel decision-making, nationally based norms are a fairer basis for comparison than local (institutional or departmental) norms.

35. A good way to understand both the range and direction(s) of student opinion in a class is to examine the percentage of students responding to each option for an item.

36. For items with a five point scale, standard deviations of less than 1.5 generally indicate that the respondents were in relatively good agreement on an item.

37. For items with a five point scale, standard deviations of more than 1.7 make the results for that item nearly useless for interpretation for any purpose.

38. For student ratings a mean score with a percentile ranking of 65 usually indicates better teaching performance than a mean with a percentile ranking of 50.

39. For an item such as “rate this instructor compared to others you have had,” a mean score with a percentile ranking of 10 indicates a poor teaching performance.

40. Student ratings usually fall within a classic “bell curve” so that the “average” rating on an item with a response scale of 1 to 5 points would be very close to 3.00.

41. When comparing an instructor’s ratings to those of other instructors, standardized scores such as z-scores or t-scores are a better measure than directly comparing item means.

INSTRUCTIONS: PART 2

For items 42-51, you are a member of department XYZ’s promotion and tenure committee reviewing the teaching effectiveness of five instructors. Many sources of evidence concerning their teaching performance have already been considered by the committee. The committee’s attention now turns to the available student ratings for the five. The ratings were obtained using a short questionnaire which included the three items below. The results are summarized in the table on the facing page.
KEY TO QUESTIONNAIRE ITEMS

STUDENT’S SELF-REPORT OF AMOUNT LEARNED

5 pts  exceptional amount
4 pts  more than usual
3 pts  about as much as usual
2 pts  less than usual
1 pt  almost nothing

STUDENT’S OVERALL RATING OF INSTRUCTOR

5 pts  among the best
4 pts  better than average
3 pts  about average
2 pts  worse than average
1 pt  among the worst

STUDENT’S OVERALL RATING OF COURSE

5 pts  among the best
4 pts  better than average
3 pts  about average
2 pts  worse than average
1 pt  among the worst

At various times during the course of deliberations, the statements at the bottom of the facing page are made by members of the committee. Based on the data in this report alone, use the response options below to indicate your agreement with each statement (i.e. items 42 to 51).

A = strongly agree
B = tend to agree
C = tend to disagree
D = strongly disagree
E = uncertain, or can’t determine based on the available ratings data.

KEYS TO RANKS:

RANK  PERCENTILE GROUP
HI  top 10%  (91-100)
HM  next upper 20%  (71-90)
MI  middle 40%  (31-70)
LM  next lower 20%  (11-30)
LO  bottom 10%  (1-10)

STATISTICS:

MEAN for each course is the mean of student responses for each item.

MEDIAN for each course is the median of student responses.

"t" is T score for each mean compared to department XYZ sample.

RANK is percentile group in which mean fall when all means in XYZ sample are rank ordered (see key below).

REPORT OF STUDENT RATINGS OF INSTRUCTION IN THE XYZ DEPARTMENT

The average ratings in the XYZ department
(308 courses in sample) are:

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean*</th>
<th>SD</th>
<th>Median</th>
<th>High</th>
<th>Low**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learned</td>
<td>3.53</td>
<td>0.5</td>
<td>3.58</td>
<td>4.80</td>
<td>2.26</td>
</tr>
<tr>
<td>Instructor</td>
<td>3.60</td>
<td>0.7</td>
<td>3.67</td>
<td>5.00</td>
<td>1.50</td>
</tr>
<tr>
<td>Course</td>
<td>3.47</td>
<td>0.5</td>
<td>3.53</td>
<td>4.68</td>
<td>1.89</td>
</tr>
</tbody>
</table>

* mean of section means  ** highest and lowest section mean in sample

The average ratings in the university
(4069 courses in sample) are:

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>High</th>
<th>Low *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learned</td>
<td>3.63</td>
<td>0.5</td>
<td>3.67</td>
<td>5.00</td>
<td>1.33</td>
</tr>
<tr>
<td>Instructor</td>
<td>3.86</td>
<td>0.7</td>
<td>3.86</td>
<td>5.00</td>
<td>1.38</td>
</tr>
<tr>
<td>Course</td>
<td>3.67</td>
<td>0.6</td>
<td>3.73</td>
<td>5.00</td>
<td>1.56</td>
</tr>
</tbody>
</table>

PROFESSOR SMITH

<table>
<thead>
<tr>
<th>COURSE</th>
<th>LEARNED</th>
<th>INSTRUCTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>XYZ 101-01</td>
<td>3.78</td>
<td>0.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COURSE</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.63</td>
<td>0.7</td>
<td>53</td>
<td>MI</td>
<td>43/40              (93%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COURSE</th>
<th>LEARNED</th>
<th>INSTRUCTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>XYZ 101-03</td>
<td>3.43</td>
<td>0.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COURSE</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.32</td>
<td>1.1</td>
<td>45</td>
<td>MI</td>
<td>30/28              (93%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COURSE</th>
<th>LEARNED</th>
<th>INSTRUCTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>XYZ 101-02</td>
<td>3.29</td>
<td>0.7</td>
</tr>
</tbody>
</table>

<p>| COURSE |</p>
<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.41</td>
<td>0.9</td>
<td>49</td>
<td>MI</td>
<td>42/34 (81%)</td>
</tr>
</tbody>
</table>

**PROFESSOR ASPEN**

<table>
<thead>
<tr>
<th>COURSE</th>
<th>LEARNED</th>
<th>INSTRUCTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>T</td>
</tr>
<tr>
<td>XYZ 270-01</td>
<td>3.67</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**COURSE**

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.17</td>
<td>0.4</td>
<td>64</td>
<td>HI</td>
<td>7/6 (86%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COURSE</th>
<th>LEARNED</th>
<th>INSTRUCTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>T</td>
</tr>
<tr>
<td>XYZ 402-01</td>
<td>4.20</td>
<td>0.7</td>
</tr>
</tbody>
</table>

**COURSE**

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.60</td>
<td>0.5</td>
<td>53</td>
<td>HI</td>
<td>5/5 (100%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COURSE</th>
<th>LEARNED</th>
<th>INSTRUCTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>T</td>
</tr>
<tr>
<td>XYZ 401-01</td>
<td>4.17</td>
<td>0.7</td>
</tr>
</tbody>
</table>

**COURSE**

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.17</td>
<td>0.7</td>
<td>64</td>
<td>HI</td>
<td>7/6 (86%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COURSE</th>
<th>LEARNED</th>
<th>INSTRUCTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>T</td>
</tr>
<tr>
<td>XYZ 401-02</td>
<td>4.20</td>
<td>0.7</td>
</tr>
</tbody>
</table>

**COURSE**

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.29</td>
<td>0.7</td>
<td>64</td>
<td>HI</td>
<td>5/5 (100%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COURSE</th>
<th>LEARNED</th>
<th>INSTRUCTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>T</td>
</tr>
<tr>
<td>XYZ 402-01</td>
<td>4.80</td>
<td>0.4</td>
</tr>
</tbody>
</table>

**COURSE**

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.60</td>
<td>0.5</td>
<td>70</td>
<td>HI</td>
<td>5/5 (100%)</td>
</tr>
</tbody>
</table>

**PROFESSOR FAHEY**

<table>
<thead>
<tr>
<th>COURSE</th>
<th>LEARNED</th>
<th>INSTRUCTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>T</td>
</tr>
<tr>
<td>XYZ 401-01</td>
<td>3.30</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**COURSE**

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LEARNED COURSE</td>
<td>Mean</td>
<td>SD</td>
<td>T</td>
</tr>
<tr>
<td>------------</td>
<td>----------------</td>
<td>------</td>
<td>-----</td>
<td>---</td>
</tr>
<tr>
<td>XYZ 401-01</td>
<td>3.30</td>
<td>0.9</td>
<td>49</td>
<td>MI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>COURSE</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYZ 401-01</td>
<td>3.35</td>
<td>1.1</td>
<td>46</td>
<td>MI</td>
<td>2.76</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>2.85</td>
<td>0.9</td>
<td>38</td>
<td>LM</td>
<td>61/29</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>COURSE</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYZ 401-01</td>
<td>3.97</td>
<td>1.4</td>
<td>59</td>
<td>HM</td>
<td>3.57</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>3.63</td>
<td>0.7</td>
<td>53</td>
<td>MI</td>
<td>39/16</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>LEARNED COURSE</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>LEARNED INSTRUCTOR</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYZ 402-01</td>
<td>3.30</td>
<td>1.0</td>
<td>45</td>
<td>LM</td>
<td>3.00</td>
<td>1.0</td>
<td>42</td>
<td>LM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>COURSE</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYZ 402-01</td>
<td>3.83</td>
<td>0.8</td>
<td>57</td>
<td>HM</td>
<td>3.67</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>3.50</td>
<td>0.6</td>
<td>51</td>
<td>MI</td>
<td>14/12</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>LEARNED COURSE</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>LEARNED INSTRUCTOR</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYZ 402-01</td>
<td>3.57</td>
<td>0.8</td>
<td>51</td>
<td>MI</td>
<td>3.43</td>
<td>1.1</td>
<td>48</td>
<td>MI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>COURSE</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYZ 402-01</td>
<td>4.33</td>
<td>0.6</td>
<td>67</td>
<td>HI</td>
<td>4.08</td>
<td>0.8</td>
</tr>
<tr>
<td>Mean</td>
<td>SD</td>
<td>T</td>
<td>Rank</td>
<td>Enrolled/Responded</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>--------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.17</td>
<td>0.7</td>
<td>64</td>
<td>HI</td>
<td>16/12 (75%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**LEARNED COURSE**

<table>
<thead>
<tr>
<th>XYZ 402-01</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.55</td>
<td>0.7</td>
<td>50</td>
<td>MI</td>
</tr>
</tbody>
</table>

**INSTRUCTOR**

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.60</td>
<td>0.8</td>
<td>57</td>
<td>HM</td>
</tr>
</tbody>
</table>

**Professor Parker**

<table>
<thead>
<tr>
<th>LEARNED COURSE</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYZ 402-01</td>
<td>3.49</td>
<td>0.9</td>
<td>49</td>
<td>MI</td>
</tr>
</tbody>
</table>

**INSTRUCTOR**

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.32</td>
<td>0.9</td>
<td>60</td>
<td>HM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COURSE</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.65</td>
<td>0.8</td>
<td>54</td>
<td>MI</td>
<td>39/37 (95%)</td>
</tr>
</tbody>
</table>

**LEARNED COURSE**

<table>
<thead>
<tr>
<th>XYZ 402-01</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.82</td>
<td>0.8</td>
<td>56</td>
<td>HM</td>
</tr>
</tbody>
</table>

**INSTRUCTOR**

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.09</td>
<td>0.9</td>
<td>57</td>
<td>HM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COURSE</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Rank</th>
<th>Enrolled/Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.86</td>
<td>1.0</td>
<td>58</td>
<td>HM</td>
<td>33/22 (67%)</td>
</tr>
</tbody>
</table>

42. Aspen is more effective than Smith.
43. Smith is more effective than Fahey.
44. Cohen is less effective than Fahey.
45. Cohen is less effective than Aspen.
46. Parker is more effective than Fahey or Cohen.
47. Fahey is the least effective of the group.
48. If only one of these instructors were to receive a merit raise, Aspen is the best choice of the five.
49. It is not fair to compare Smith and Cohen with each other.
50. Instructors in XYZ department are rated relatively lower than instructors in the university at large.
51. With the possible exception of Smith and Cohen, it is impossible to fairly compare the overall teaching effectiveness of any two instructors within this group, based on the ratings presented in this report.

**INSTRUCTIONS: PART 3**

Use the response scale below for the items 52-63.

A = Strongly Agree
B = Agree  
C = Slightly Agree  
D = Slightly Disagree  
E = Disagree  
F = Strongly Disagree

52. Student rating questionnaires are not a very useful way to measure teaching effectiveness in my particular discipline.

53. I know enough about statistics to interpret the results of student ratings without assistance.

54. I don't really know how to use the results of student ratings to improve my teaching skills or course.

55. Administrators generally use student ratings in promotion and tenure decisions in ways that accurately depict quality of teaching performance.

56. Students have a right/responsibility to rate courses/instructors.

57. The use of student ratings in my institution provides little or no benefit to the quality of instruction students receive.

58. I believe mandatory evaluation of teaching performance without institutional support for teaching improvement is punitive.

59. I trust my students to give sincere and honest responses on student rating questionnaires.

60. I have been able to improve my teaching or my course based on information I obtained from student ratings.

61. With the exception of notably poor teaching performance, there appears to be little or no practical administrative interest in the quality of teaching in my department.

62. I am not aware of any explicit procedures or policies regarding the use of student ratings in personnel decisions in my department.

63. I feel my career has been harmed to some degree by student ratings I have received.

64. Indicate the response which best fits your current duties (excluding research).  
A = teaching only  
B = teaching with administrative duties, e.g. department chair  
C = primarily administrative appointment with some teaching duties  
D = administrative appointment, no teaching duties  
E = instructional development, faculty development, or evaluation professional, with some teaching duties  
F = instructional development, faculty development, or evaluation professional, no teaching duties

65. How many years have you been teaching at the post-secondary level?  
A = less than one  
B = one to two  
C = more than two but less than eight  
D = eight or more but less than twelve  
E = more than twelve but less than twenty  
F = twenty or more
66. Indicate your faculty rank. (Leave blank if non-faculty appointment).
A = full professor (including emeritus)
B = associate professor
C = assistant professor
D = instructor
E = lecturer (including adjunct, senior, and part-time
F = teaching assistant
67. Indicate how often you have participated in personnel decision-making in
which student ratings (other than your own) were offered as evidence of
教学 performance.
A = on a regular basis (at least once every academic year for at least the last two
years).
B = not currently, but have done so at least twice in my career
C = have never done so
68. Indicate which response describes how you have participated in created or
selecting student ratings forms.
A = I have chose forms for use in my department or institution for personnel
decision-making.
B = I have helped write or adapt forms for use in my department or institution for
personnel decision-making.
C = I chosen, adapted, or written forms for use by others for use in my department
or institution for diagnosis or teaching improvement.
D = I have chosen or written forms for my own use to present evidence of my
teaching performance for personnel decision-making.
E = I have chosen or written forms for my own use in obtaining diagnostic
feedback for teaching improvement.
F = I have never selected, adapted, or written such forms.
69. Have you ever received assistance from a teaching improvement specialist or
master teacher in your efforts to interpret and/or use student ratings to
improve the instruction you provide?
A = YES, on at least two separate occasions
B = YES, but only once
C = NO, never, BUT I would to try it sometime
D = NO, never, AND I am not interested in trying it
Appendix J

Expert Group for 1989 Survey
**Expert Researchers/Writers/Practitioners Who Participated in the First Validation of the Survey Item Bank**

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Philip Abrami</td>
<td>Concordia University</td>
</tr>
<tr>
<td>Dr. Lawrence Aleamoni</td>
<td>Arizona State University</td>
</tr>
<tr>
<td>Dr. Raoul Arreola</td>
<td>University of Tennessee</td>
</tr>
<tr>
<td>Dr. Dale Brandenburg</td>
<td>University of Illinois</td>
</tr>
<tr>
<td>Dr. Mary Ann Bunda</td>
<td>Western Michigan University</td>
</tr>
<tr>
<td>Dr. William Cashin</td>
<td>Kansas State University</td>
</tr>
<tr>
<td>Dr. John Centra</td>
<td>Syracuse University</td>
</tr>
<tr>
<td>Dr. Peter Cohen</td>
<td>Medical College of Georgia</td>
</tr>
<tr>
<td>Dr. Patricia Cranton</td>
<td>Brock University</td>
</tr>
<tr>
<td>Dr. Kenneth Doyle</td>
<td>University of Minnesota</td>
</tr>
<tr>
<td>Dr. Glenn Erickson</td>
<td>University of Rhode Island</td>
</tr>
<tr>
<td>Dr. Stanford Erickson</td>
<td>University of Florida</td>
</tr>
<tr>
<td>Dr. Kenneth Feldman</td>
<td>SUNY Stonybrook</td>
</tr>
<tr>
<td>Dr. Peter Frey</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>Dr. George Geis</td>
<td>Ontario Institute for Studies in Higher Education</td>
</tr>
<tr>
<td>Dr. Gerry Gillmore</td>
<td>University of Washington</td>
</tr>
<tr>
<td>Dr. Christopher Knapper</td>
<td>Waterloo University</td>
</tr>
<tr>
<td>Dr. James Kulik</td>
<td>University of Michigan</td>
</tr>
<tr>
<td>Dr. Wilbert McKeachie</td>
<td>University of Michigan</td>
</tr>
<tr>
<td>Dr. Phillip McKnight</td>
<td>University of Kansas</td>
</tr>
<tr>
<td>Dr. Robert Menges</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>Dr. Harry Murray</td>
<td>University of Western Ontario</td>
</tr>
<tr>
<td>Dr. Peter Seldin</td>
<td>Pace University</td>
</tr>
<tr>
<td>Dr. Mary Deane Sorcinelli</td>
<td>Indiana University</td>
</tr>
</tbody>
</table>
Appendix K

Instrument Used in this Study
Survey on Student Course Evaluations

1. Please circle your faculty rank (leave blank if non-faculty appointment).
   A = Full Professor    B = Associate Professor    C = Assistant Professor
   D = Instructor E = Adjunct, Part-Time    F = Teaching Assistant

2. Please circle the number of years you have worked in higher education.
   A = Less than 3    B = 3 to 6 years    C = 6 to 9 years
   D = 9 to 12 years    E = 12 to 15 years    F = more than 15

3. Please circle the response that best fits your current duties.
   A = teaching only    B = teaching w/administrative duties
   C = administrative only    D = administrative w/ teaching duties

4. Please circle how often you have participated in personnel decisions in which
   student ratings were offered as evidence of teaching performance.
   A = Often (at least once in the past two years) B = Sometimes (at least once in the
   past four years) C = Never

5. Please circle the response that best describes your level of participation in creating
   or selecting student ratings forms.
   A = I have helped create or adapt student evaluation forms.
   B = I have help to choose student evaluation forms.
   C = I have never created or adapted student ratings forms.
   D = I have never helped to choose student ratings forms.

For the remainder of the survey (questions 6-25) please circle the response that best
indicates your reaction.

6. There is rarely good agreement between the ratings of students and other
   measures of teaching competence (such as administrative and peer reviews).
   Strongly Agree    Tend to Agree    Tend to Disagree    Strongly Disagree    Uncertain

7. Student ratings questionnaires properly constructed and administered can do a
   good enough job to warrant their use for personnel decisions.
   Strongly Agree    Tend to Agree    Tend to Disagree    Strongly Disagree    Uncertain

8. Student ratings can't really give a teacher the kind of feedback needed for
   improving the quality of instruction s/he provides.
   Strongly Agree    Tend to Agree    Tend to Disagree    Strongly Disagree    Uncertain

9. The ratings obtained from "good" students should be given more weight than
   those of students who are not doing well (i.e. the higher the grade point average of
   the rater, the more valid is his or her response).
   Strongly Agree    Tend to Agree    Tend to Disagree    Strongly Disagree    Uncertain

10. The proportion of a class that rates an instructor is not as important as the total
    number of raters.
   Strongly Agree    Tend to Agree    Tend to Disagree    Strongly Disagree    Uncertain
11. Very specific items dealing with a teacher's behavior and skills are usually more reliable than broad, global items such as "this instructor compared to others."

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
</table>

12. Even a few written comments from students always tell more about a course than a full page of statistics from a student ratings questionnaire, no matter how many students respond to it.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
</table>

13. Discussion, lecture, and laboratory classes are so different that there is no valid way to only use student ratings to compare the instructors who teach them.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
</table>

14. When ratings are mandatory, every course/section of every instructor should be evaluated every academic term.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
</table>

15. Generally, a rating of heavier than average workload and/or difficulty is associated with lower than average ratings of the teacher's performance based on "overall" item scores.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
</table>

16. For items on a five point scale, standard deviations of less than 1.2 generally indicate that the respondents were in relatively good agreement on an item.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
</table>

17. For an item such as "rate this instructor compared to others you have had," a mean score with a percentile ranking in the lowest 10% indicates a poor teaching performance.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
</table>

18. Student ratings usually fall within a classic "bell curve" so that the "average" rating on an item with a response scale of 1 to 5 points would be very close to 3.00.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
</table>

19. When comparing an instructor's ratings to those of other instructors, standardized scores such as z-scores or t-scores provide a more accurate measure than direct comparisons of item means.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
</table>

20. A good way to understand both the range and direction(s) of student opinion in a class is to examine the percentages of students responding to each option for an item.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
</table>

21. Student rating questionnaires are not a very useful way to measure teaching effectiveness.
22. Students have a right/responsibility to rate courses/instructors.

Strongly Agree  Tend to Agree  Tend to Disagree  Strongly Disagree  Uncertain

23. Overall items in student course evaluations provides little or no information for course improvement.

Strongly Agree  Tend to Agree  Tend to Disagree  Strongly Disagree  Uncertain

24. Students generally provide sincere, honest and accurate responses to questions about teaching effectiveness.

Strongly Agree  Tend to Agree  Tend to Disagree  Strongly Disagree  Uncertain

25. Specific questions about teaching behaviors in student course evaluations can help me improve my teaching.

Strongly Agree  Tend to Agree  Tend to Disagree  Strongly Disagree  Uncertain

For Coding Purposes Only:

YEAR OF BIRTH:  MOTHER'S MAIDEN NAME:
Appendix L

Numeric Equivalents of 6-25 on Instrument
Survey on Student Course Evaluations  
Numeric Equivalents of Questions 6-25

1. Please circle your faculty rank (leave blank if non-faculty appointment).  
A = Full Professor    B = Associate Professor    C = Assistant Professor  
D = Instructor        E = Adjunct, Part-Time       F = Teaching Assistant

2. Please circle the number of years you have worked in higher education.  
A = Less than 3       B = 3 to 6 years           C = 6 to 9 years       
D = 9 to 12 years     E = 12 to 15 years        F = more than 15

3. Please circle the response that best fits your current duties.  
A = teaching only     B = teaching w/administrative duties  
C = administrative only   D = administrative w/ teaching duties

4. Please circle how often you have participated in personnel decisions in which student ratings were offered as evidence of teaching performance.  
A = Often (at least once in the past two years)  B = Sometimes (at least once in the past four years)  C = Never

5. Please circle the response that best describes your level of participation in creating or selecting student ratings forms.  
A = I have helped create or adapt student evaluation forms.  
B = I have help to choose student evaluation forms.  
C = I have never created or adapted student ratings forms.  
D = I have never helped to choose student ratings forms. 

For the remainder of the survey (questions 6-25) please circle the response that best indicates your reaction.

6. There is rarely good agreement between the ratings of students and other measures of teaching competence (such as administrative and peer reviews).  

7. Student ratings questionnaires properly constructed and administered can do a good enough job to warrant their use for personnel decisions.  

8. Student ratings can't really give a teacher the kind of feedback needed for improving the quality of instruction s/he provides.  

9. The ratings obtained from "good" students should be given more weight than those of students who are not doing well (i.e. the higher the grade point average of the rater, the more valid is his or her response).
10. The proportion of a class that rates an instructor is not as important as the total number of raters.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

11. Very specific items dealing with a teacher’s behavior and skills are usually more reliable than broad, global items such as "this instructor compared to others."

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

12. Even a few written comments from students always tell more about a course than a full page of statistics from a student ratings questionnaire, no matter how many students respond to it.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

13. Discussion, lecture, and laboratory classes are so different that there is no valid way to only use student ratings to compare the instructors who teach them.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

14. When ratings are mandatory, every course/section of every instructor should be evaluated every academic term.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

15. Generally, a rating of heavier than average workload and/or difficulty is associated with lower than average ratings of the teacher’s performance based on "overall" item scores.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

16. For items on a five point scale, standard deviations of less than 1.2 generally indicate that the respondents were in relatively good agreement on an item.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

17. For an item such as "rate this instructor compared to others you have had," a mean score with a percentile ranking in the lowest 10% indicates a poor teaching performance.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

18. Student ratings usually fall within a classic "bell curve" so that the "average" rating on an item with a response scale of 1 to 5 points would be very close to 3.00.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

19. When comparing an instructor's ratings to those of other instructors, standardized scores such as z-scores or t-scores provide a more accurate measure than direct comparisons of item means.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
20. A good way to understand both the range and direction(s) of student opinion in a class is to examine the percentages of students responding to each option for an item.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

21. Student rating questionnaires are not a very useful way to measure teaching effectiveness.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

22. Students have a right/responsibility to rate courses/instructors.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

23. Overall items in student course evaluations provides little or no information for course improvement.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

24. Students generally provide sincere, honest and accurate responses to questions about teaching effectiveness.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

25. Specific questions about teaching behaviors in student course evaluations can help me improve my teaching.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Strongly Agree</th>
<th>Tend to Agree</th>
<th>Tend to Disagree</th>
<th>Strongly Disagree</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

For Coding Purposes Only:

YEAR OF BIRTH:  
MOTHER'S MAIDEN NAME: