University of San Diego

Digital USD

Dissertations

Theses and Dissertations

2017

An Empirical Look at the Impact of Course and Faculty Characteristics on Student Evaluations

Travis J. Degheri
University of San Diego

Follow this and additional works at: <https://digital.sandiego.edu/dissertations>

Part of the [Higher Education Commons](#)

Digital USD Citation

Degheri, Travis J., "An Empirical Look at the Impact of Course and Faculty Characteristics on Student Evaluations" (2017). *Dissertations*. 76.

<https://digital.sandiego.edu/dissertations/76>

This Dissertation: Open Access is brought to you for free and open access by the Theses and Dissertations at Digital USD. It has been accepted for inclusion in Dissertations by an authorized administrator of Digital USD. For more information, please contact digital@sandiego.edu.

AN EMPIRICAL LOOK AT THE IMPACT OF COURSE AND FACULTY
CHARACTERISTICS ON STUDENT EVALUATIONS

by

Travis J. Degheri

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

May 2017

Dissertation Committee

Fred J. Galloway, Ed.D.
Christopher B. Newman, Ph.D.
Paula S. Krist, Ph.D.
Iris H. W. Engstrand, Ph.D.
Pat Burk, Ph.D.
Patricia Heydet-Kirsch, Ed.D.

University of San Diego

ABSTRACT

Student evaluations of college instructors are anything but a new phenomenon, having been used since the early 1900s. Today, universities around the world continue to use student evaluations as a means for measuring instructor effectiveness. Despite concerns of student objectivity, at many institutions these evaluations have a sizable influence on decisions involving faculty promotion, tenure, and merit salary increases. While there is much literature examining student evaluations, few studies have provided a longitudinal, multi-discipline exploration of the impact course and faculty characteristics have on student evaluations.

To address this gap in the literature, this study used publically available data collected over two consecutive academic years from a single college located within a university in the southeastern United States to examine the extent to which course and faculty characteristics explained variation in undergraduate student evaluations. Mean and median scores associated with quality of instruction, amount of student learning, and relative performance of the professor were used as dependent variables in the analysis of 1,812 separate classes. Findings revealed that select course and faculty characteristics explained a significant amount of the variation in student evaluations. For example upper division courses tended to receive better ratings than lower division courses; early morning courses received lower ratings than any other time; general education courses tended to be scored lower than non-general education course; and more often than not, female professors received lower ratings than their male counterparts. Scores also varied significantly by department as did the patterns of significance among the more than 40 independent variables used in the analysis. Interestingly, political science and history had

the highest scores and anthropology and philosophy the lowest. Taken together, these models explained between 1% and 45% of the variation in evaluation scores among the 11 departments used in the analysis.

Given the important role that student evaluations play in the decision-making process underlying faculty promotion, tenure, and merit salary increases, the findings in this study will help both faculty and administrators better understand the course and instructor characteristics that may be impacting student evaluations, in effect creating a more equitable and efficient process for reviewing faculty.

ACKNOWLEDGEMENTS

I have been so fortunate to have had numerous people support me throughout this dissertation process and graduate school in general. I would first like to thank Dr. Fred Galloway, my committee chair. You are much more than an advisor or instructor to me; you are a friend and I feel so blessed to have had your encouragement and friendship throughout this program. I am also so very grateful for the five other committee members, Dr. Christopher Newman, Dr. Paula Krist, Dr. Iris Engstrand, Dr. Pat Burk, and Dr. Patricia Heydet-Kirsch, all of whom worked so well together. While a committee of six may have initially seemed daunting, your perspectives, edits, support, and friendship has been invaluable and, now, I cannot envision going through this process any other way.

I have spent the past twelve years at the University of San Diego for my undergraduate, master's and now doctoral degree. It has been like a home to me, not only for the time I've spent there, but for the people I met there that have become like family. In addition to those already mentioned, I must give many thanks to others who have enriched my experience at USD including, but not limited to: Dr. Timothy O'Malley, Dr. Mary Lyons, Dr. Robert Donmoyer, Dr. Kristin Moran, Dr. Reyes Quezada, Dr. Mark Peters, Dr. Rafael Tovar Y Lopez-Portillo, Dr. Nicholas Franco, Dr. Pressley Rob Rankin, Linda Dews, Rondi Stein, Beth Garofalo, Milton Maples, Simone Blancato, Kevin Holt, and Raymundo Quezada. I am thankful for all this university has given me and all the people that make USD a place to call home.

I would also like to thank my siblings and especially my parents, who have always gone above and beyond to support me in my academic endeavors. Finally, I

would like to thank my wife, Sarah Degheri, who I met during my freshman year at USD.

I admire and thank you for your strength and patience you've displayed over the years.

Thank you for being the wonderful wife, mother, and best friend that you are. I love you.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	iv
TABLES OF CONTENTS.....	vi
LIST OF TABLES.....	ix
CHAPTER ONE: BACKGROUND.....	1
Problem Statement.....	3
Purpose of the Study and Research Questions.....	3
CHAPTER TWO: REVIEW OF THE LITERATURE.....	5
Introduction.....	5
Faculty Characteristics.....	7
Gender.....	7
Race/Ethnicity.....	8
Tenure.....	10
Personality.....	11
Course Characteristics.....	11
Academic Discipline.....	12
Online Versus Onsite.....	12
Class Time.....	14
Class Size.....	14
Course Level.....	15
Student Characteristics.....	16
Age.....	16
Gender.....	17

Anticipated or Actual Grade.....	17
Conclusion.....	19
CHAPTER THREE: METHODOLOGY.....	20
Research Site and Datasets.....	20
Institutional Data.....	21
Student demographics.....	22
Faculty demographics.....	22
Course characteristics.....	22
Datasets.....	23
Student Perceptions of Teaching (SPOT) Instrument.....	25
Course Schedule.....	27
Additional Resources.....	28
CHAPTER FOUR: ANALYSIS.....	30
Sample Characteristics.....	30
Research Question 1.....	31
Research Question 2.....	37
CHAPTER FIVE: DISCUSSION.....	57
Purpose of the Study.....	57
Methodology Review.....	57
Discussion of Findings.....	59
Question 1.....	59
Question 2.....	64
Faculty characteristics.....	64

Course characteristics.....	70
Course and faculty characteristics.....	74
The relationship of course and faculty characteristics with individual questions and departments.....	83
Discussion Summary.....	85
Implications for Policy.....	86
Limitations.....	88
Implications for Future Research.....	89
REFERENCES.....	92
APPENDIX: SAMPLE SPOT SURVEY.....	96

LIST OF TABLES

Table 1. Total Courses Per Department.....	31
Table 2. Descriptive Statistics for All Observations (All Departments).....	32
Table 3. SPOT Database Question 19 Example.....	33
Table 4. SPOT Database Question 20 Example.....	33
Table 5. SPOT Database Question 21 Example.....	34
Table 6. Means and Standard Deviations by Department for the Three SPOT Questions.....	35
Table 7. Medians and Standard Deviations by Department for the Three SPOT Questions.....	35
Table 8. One-Way ANOVA for the Three SPOT Questions Based on Department.....	36
Table 9. Kruskal-Wallis Test of Medians for the Three SPOT Questions by Department.....	36
Table 10. Faculty Characteristics Regression Based on Means of the Three SPOT Questions by Department.....	40
Table 11. Faculty Characteristics Regression Based on Medians of the Three SPOT Questions by Department.....	41
Table 12. Faculty Characteristics Correlations.....	42
Table 13. Estimated Coefficients and Levels of Significance for Course Characteristics Regression Based on Mean from the Student Learning Question.....	44
Table 14. Estimated Coefficients and Levels of Significance for Course Characteristics Regression Based on Mean from the Instructional Quality Question.....	45

Table 15. Estimated Coefficients and Levels of Significance for Course Characteristics	
Regression Based on Mean from the Instructor Comparison Question.....	46
Table 16. Estimated Coefficients and Levels of Significance for Course Characteristics	
Regression Based on Median from the Student Learning Question.....	47
Table 17. Estimated Coefficients and Levels of Significance for Course Characteristics	
Regression Based on Median from the Instructional Quality Question.....	48
Table 18. Estimated Coefficients and Levels of Significance for Course Characteristics	
Regression Based on Median from the Instructor Comparison Question.....	49
Table 19. Regression for All Course and Faculty Characteristics Stepwise for	
Anthropology.....	50
Table 20. Regression for All Course and Faculty Characteristics Stepwise for Art.....	51
Table 21. Regression for All Course and Faculty Characteristics Stepwise for	
Communications.....	51
Table 22. Regression for All Course and Faculty Characteristics Stepwise for English...	51
Table 23. Regression for All Course and Faculty Characteristics Stepwise for History...	52
Table 24. Regression for All Course and Faculty Characteristics Stepwise for	
Languages.....	52
Table 25. Regression for All Course and Faculty Characteristics Stepwise for Music.....	52
Table 26. Regression for All Course and Faculty Characteristics Stepwise for	
Philosophy.....	53
Table 27. Regression for All Course and Faculty Characteristics Stepwise for Political	
Science.....	53

Table 28. Regression for All Course and Faculty Characteristics Stepwise for	
Sociology.....	53
Table 29. Regression for All Course and Faculty Characteristics Stepwise for Theatre and	
Dance.....	54
Table 30. Summarized Regression for All Course and Faculty Characteristics for All	
Departments.....	55

CHAPTER ONE

BACKGROUND AND PURPOSE OF THE STUDY

Student evaluations¹ are anything but a new phenomenon, having been used in colleges and universities since the early 1900s. Over time, these evaluations have evolved from in-class evaluations via paper and pencil instruments to online evaluations that can be completed at the student's convenience. This technology, however, has its limitations. For one, the university is relying on individuals to complete these evaluations outside of class and usually after the course is completed. Institutions do not typically require participation and there has been a noticeable decline in online response rates compared to the traditional paper and pencil evaluations (Stowell, Addison, & Smith, 2012; Spooren, Brockx, & Mortelmans, 2013), though some scholars have found online evaluations to yield a higher response rate (Venette, Sellnow, & McIntyre, 2010).

Online evaluations are here to stay; they are the wave of the future. Online evaluations have gained popularity from economic, accessibility and environmental perspectives. For example, online evaluations are available to faculty and administrators conveniently and, in some cases, are made available to students and the general public. In fact, the university in this study was partly chosen for its accessibility, as their student evaluations are publicly available online.

Nevertheless, there is the concern that if online evaluations are to be at least as effective as paper and pencil evaluations, they should have comparable response rates. Despite lower response rates, online evaluations actually tend to generate more data. In

¹ Over time, the terminology in examining student evaluations has changed. For the sake of consistency in this study, unless otherwise noted, I will use "student evaluations" to describe any case where students rate the faculty and/or course quantitatively.

terms of open-ended questions on these evaluations, students who complete the online evaluations tend to give more thorough and thoughtful responses (Venette, et al., 2010), which are valued by many as more insightful than the responses to the Likert-scale questions on the evaluations. Additionally, online evaluation tools make it easier to organize collected data and examine changes in student responses for a particular faculty member or course over time, since most analysis is already completed via computers.

As evidenced by the literature, student evaluations have become widely used as a means for measuring instructor effectiveness (Marsh, 1984). However, many researchers remain unconvinced that they are an adequate or unbiased measure of true teaching effectiveness, which is problematic when student evaluations are used in decisions regarding promotion, tenure, or even salary increases (McPherson & Jewell, 2007; Zhao & Gallant, 2012; Harris & Twinaime, 2012; Cisneros-Cohernour, 2005; Arbuckle & Williams, 2003, Young, 1993).

Not surprisingly, one of the biggest complaints from both faculty and administrators regarding student evaluations of instructors is their potential subjectivity (Xu, 2012; Shirbagi, 2011; Spooren, Brockx, & Mortelmans, 2013; Young, 1993; Cisneros-Cohernour, 2005). There are conflicting findings in the literature regarding the objectivity of students, but these studies are typically hampered by methodological issues that limit the size or scope of student responses. For example, McPherson & Jewell (2007) conducted one of the few studies to extensively examine items separate from the student evaluations, such as, gender, race, tenure, and class size. This study had limitations in that it only looked at one program at one school. Still, this study suggests that improving the characteristics of sample sizes and instructors over a broader range of

courses may lead to better understanding of the possible linkages between the characteristics of the course and instructor and the results of the student evaluations.

Problem Statement

There is little evidence in the literature that describes the contextual factors that might impact student evaluations of instructors. Researchers have not thoroughly explored course and faculty characteristics that may directly impact student evaluations. In addition, many of these studies are limited in either size or scope, as they just focus on a limited number of courses or only stay within one discipline. This study addresses the fundamental problems of both whether, and how, student course evaluations are impacted by contextual variables of the course.

In an effort to fill in these gaps in the literature, this study will examine two aspects of contextual variance in particular. The first is the extent to which student perceptions of instructors vary both within and across disciplines. The second is the extent to which instructor and course characteristics are associated with variation in student evaluations. By responding to these two points, this study will help fill an identified gap in the literature, and encourage other scholars to take on similar studies.

Purpose of the Study and Research Questions

As mentioned in the preceding section, this study is designed to examine two things in particular: the extent to which student evaluations vary both within and across disciplines, and the extent to which instructor and course characteristics are associated with variation in student ratings. These will be addressed in the following two research questions:

1. To what extent do student evaluations vary by academic discipline?

2. To what extent can instructor and course characteristics explain variation in student evaluations?

CHAPTER TWO

REVIEW OF THE LITERATURE

Introduction

Student evaluations are not a new phenomenon, as evidenced by the extensive literature on the subject. What was once a Western concept for evaluating teaching effectiveness has now become a standard practice in many countries. Despite its widespread influence, however, questions arise regarding the method of evaluation and the meaning of the results. In particular, are results an objective depiction of faculty skill, or are results influenced by characteristics of the students, the faculty member and the course itself that define the context within which the course exists? How has the onset of online course evaluation methods impacted the faculty and course evaluation process? The research literature on student evaluations provides insights into these issues and to various approaches to their investigation.

The use of student evaluations in higher education dates back to the early 1900s with scholars studying various factors that may affect ratings since the 1920s (Wachtel, 1998). It's unclear what the original purpose of student evaluations was, but providing students an opportunity to provide opinions on faculty reflected the early 20th century characteristics of intellectualism, individuality and equality associated with the period following World War I. This idealized conception of student and faculty relationships continued largely unchanged through the Depression and World War II. However, the emergence of both the civil rights era and anti-war movement surrounding the Vietnam War dramatically changed the nature of student voice relative to their teachers. Students demanded more voice in the affairs of their universities, including a voice on the quality

of their instructors (Murray, 2005). One of the more lasting changes from this era is the process of including student voices in the evaluation of their teachers. As the process of data collection of student evaluations became more routine on campuses around the country, more attention has been focused on the goals and purposes of student evaluations, collection procedures, and more recently, course, faculty, and student characteristics that may impact ratings.

While some faculty protest the validity of student evaluations, numerous administrators protest it as well because there is tremendous weight given to these evaluations (Xu, 2012; Shirbagi, 2011; Spooren, Brockx, & Mortelmans, 2013; Young, 1993; Cisneros-Cohernour, 2005). Though some scholars support student evaluations (Cashin, 1995; Marsh, 1984; Jirovec, Ramanathan, & Alvarez, 1998; Zhao & Gallant, 2012), many scholars think it is unwise to give the evaluations so much consideration, namely because of student biases, but also course and faculty characteristics that may impact instructor ratings on these evaluations (Cisneros-Cohernour, 2005; Sprinkle, 2008; Arbuckle & Williams, 2003; Platt, 1993; Galbraith, Merrill, & Kline, 2012).

Student evaluations have the capacity to change an individual's life, for better or worse, which is likely something students don't even consider when completing them. As noted in the literature, these evaluations are factored into tenure, promotion, and salary at many universities, not only in the United States, but also worldwide (McPherson & Jewell, 2007; Young, 1993; Zhao & Gallant, 2012; Cisneros-Cohernour, 2005; Arbuckle & Williams, 2003).

Because there is so much that depends on student evaluations and their integrity, there is a need to find an answer to the question: are student evaluations a sufficient

means for measuring teaching and/or course effectiveness? In my attempt to address this question in the literature, I will organize this review of the literature by three main categories: faculty characteristics, course characteristics, and finally, student characteristics. These three categories emerged from my exploration of the literature as having a potential impact on student evaluations.

Faculty Characteristics

For years, researchers have studied the impact faculty have on student evaluations. It is a justified curiosity, as faculty are the suppliers of content for courses and largely responsible for student learning outcomes. It's reasonable to say that the success of a course, or rather, success as perceived by students, heavily weighs on the faculty member teaching the course. But what if student learning and teaching effectiveness were afterthoughts in a student's evaluation and something entirely different, such as faculty characteristics, steered an evaluation a certain way? This section describes the most prevalent faculty characteristics discussed in the student evaluation literature.

Gender

What does a faculty member's gender have to do with student evaluations? Some researchers have found some correlation between faculty gender and student evaluations (Campbell, Gerdes, & Steiner, 2005; Wachtel, 1998; Arbuckle & Williams, 2003). Students rating female instructors lower than male instructors was the most common finding (Wachtel, 1998; Arbuckle & Williams, 2003). Other researchers argue that it is not a significant factor in student evaluations (McPherson & Jewell, 2007; Liu, 2012; Krautmann & Sander, 1999). Still, any study confirming suspicions of gender bias in

student evaluations is concerning and all the more reason to continue to address this issue in future research.

While it can be difficult to say that any one variable explains a particular rating on student evaluations, some studies deduced that a combination of variables, such as faculty gender and faculty race, might explain student evaluation ratings (Smith, 2009). These ratings could be impacted positively or negatively, depending on the correlation.

Similarly, both of these variables don't even have to be "faculty characteristics" to potentially impact student evaluations. The gender of the student and the gender of the faculty member, for example, when examined together, offer a "significant interaction" that may affect student evaluations (Bachen, McLoughlin, & Garcia, 1999, p. 193).

Student gender and its impact on student evaluations will be discussed in more detail later. Smith (2009) notes that faculty race and gender may also be dependent on course level, with undergraduate and graduate students rating faculty differently in the study. Undergraduates, for example, rated Black male faculty the lowest and graduate students rated Black females the lowest according to this study. Male faculty identified as "Other" (any other race not specified) received the highest mean scores from both undergraduate and graduate students.

Race/Ethnicity

Race has already been mentioned in relation to gender, but can race alone be a variable used to explain student evaluations? McPherson and Jewell (2007) found that, "Race appears to play a significant role in Student Evaluation of Teaching (SET) scores in our data, with white instructors earning 0.08 higher SET scores than their nonwhite

colleagues” (875).² This is a problematic finding; especially if student evaluations are considered in promotion decisions (McPherson & Jewell, 2007; Young, 1993; Zhao & Gallant, 2012; Cisneros-Cohernour, 2005; Arbuckle & Williams, 2003). Smith (2009) also found faculty race to be a factor, as previously mentioned, with Black males and Black females receiving lower ratings on student evaluations compared to their White and “Other” counterparts. While both of these studies examined typical student evaluations used at the end of the semester, Basow, Codos, and Martin (2013) offered an experimental study in which students watched a short video and subsequently take a quiz and evaluate the instructor in the video based on the material covered in the video. The instructor’s race and gender varied. “African American professors were rated higher than White professors on their hypothetical interactions with students...[but] higher [quiz] scores were obtained by students who had a White professor” (352). These hypothetical interactions referred to how engaged students were with the instructor in the video. In this study, despite reporting better rapport with African American instructors, students who had an African American instructor did not score as high on the quiz as students who had a White instructor. This is an important finding because it illustrates a discrepancy between student learning and student ratings indicating that, at least in this case, one does not necessarily inform the other.

Few studies address faculty ethnicity related to student evaluations. One of the more relevant studies, however, is Smith and Anderson’s (2005) study that looks at hypothetical course syllabi by fictional instructors with common Latino/a or Anglo

² SET (Student Evaluation of Teaching) is a common acronym in the literature to describe student evaluations. It is not the name of a particular instrument.

names. These researchers found that lenient Latino/a instructors received the highest ratings, while strict Latino/a instructors received the lowest ratings, regardless of gender. This suggests that students have preconceptions about an instructor or course from looking at the race implied on the syllabus. Though this study does not examine student evaluations in particular, it still may help inform future studies on faculty ethnicity in student evaluations by suggesting that such preconceptions may exist. What remains in question, however, and the same can be said for all course, faculty, and student characteristics discussed, are the generalizability of the findings.

Tenure

Another area of student evaluation literature that scholars are interested in is the impact of faculty tenure status on student evaluations. Though the existing research sometimes uses slightly different categories—adjunct, graduate teaching assistant, part-time, tenure-track, tenured, full-time—the overarching measurement is tenured or tenure-track versus non-tenured faculty. Though some studies found little differences in student evaluations based on the faculty member's tenure status (Thyer, Myers, & Nugent, 2011; Ronco & Cahill, 2006), others found that, “adjunct faculty receive significantly better scores than tenure-track faculty” (McPherson, & Jewell, 2007, 875). This finding might not be a complete surprise, because adjuncts are hired almost completely based on teaching ability, while for tenure-track faculty, teaching is one of several elements measured (Liu, 2012). Future studies would benefit by including the tenured/tenure track and adjunct variables to help clarify what seems to be a divided subject in the literature.

Personality

For years, scholars have argued that instructor personality impacts student evaluations (Murray, Rushton, & Paunonen, 1990; Patrick, 2011). Scholars have identified personality traits in different ways, from peer evaluation (Murray, et al., 1990) to student evaluation (Feldman, 1986). Even self-evaluation has been used to gauge instructor personality, though it is not as accurate as peer evaluation (Feldman, 1986; Kolar, Funder, & Colvin, 1996). While there have been correlations between instructor personality and student evaluations, it can be argued that it is not so much the personality that is directly impacting the evaluations, rather, the teaching behaviors an instructor takes on as a result of their inherent personalities. Erdle, Murray, and Rushton (1985) posit that it is these teaching behaviors that are being assessed differently by students, not the instructor's personality. Nevertheless, this suggests that instructor personality may explain variation in student evaluations, even if indirectly.

Given all of these faculty characteristics that may affect student evaluations, future studies would benefit by continuing to explore all faculty characteristics in relation to student evaluations.

Course Characteristics

In addition to faculty characteristics, researchers believe that there are also some course characteristics that may explain variance in student evaluations. For example, there is a significant difference in student evaluations for instructors of elective courses and instructors of general education courses, with the latter receiving lower ratings (Patrick, 2011). This, however, is just one of many course characteristics in student evaluations literature.

Academic Discipline

Similar to research regarding required courses versus elective courses, there is also research on how student evaluations vary by discipline. Humanities, art, and social sciences instructors, for example, typically receive higher ratings on student evaluations than math and science courses (Feldman, 1978; Cashin, 1995; Centra, 1993). Could this also be explained by instructor personality, or even, teaching style? Perhaps certain disciplines are more likely to use a certain style, which may explain variation in student evaluations.

Furthermore, several studies look at only one discipline (McPherson & Jewell, 2007; Basow, et al., 2013; Jirovec, et al., 1998). These are all narrowly focused studies. McPherson and Jewell (2007), for example, examined graduate economics classes at one university, over the course of about ten years. While some of the findings from these studies align with other cross-discipline studies, some do not. The studies that did not align with the rest of the literature should not be dismissed, but they should also not be taken at face value either, as the size and scope of these single-discipline studies make them less generalizable. Student evaluation research would benefit from studies examining multiple disciplines that explore both same discipline courses and courses from different disciplines.

Online Versus Onsite

There are two areas of student evaluation literature that address online versus onsite evaluations. First, were evaluations administered online or onsite in a classroom?

Second, was the course itself held online or was it held on campus, onsite? Both areas of research will be discussed in this section.

Since the early 2000s, online evaluations have gained momentum in higher education institutions (Hoffman, 2003). For the most part, online evaluation instruments and paper-and-pencil instruments produce similar results (Spooren, Brockx, & Mortelmans, 2013; Stowell, Addison, & Smith, 2012). While there are some advantages to using online student evaluations, such as more thorough comments, one disadvantage is its lower response rate compared to traditional paper-and-pencil evaluations (Spooren, et al., 2013). Venette, Sellnow, and McIntyre (2010) offer an example to the contrary, where online evaluations received higher response rates, and confirmed that open-ended comments were more thorough for online evaluations. Meanwhile, other scholars found no differences in the number of comments between online and onsite evaluations (Stowell, et al., 2012). Scholars have even gone so far as to compare RateMyProfessors.com, an example of an online public evaluation instrument, to evaluations conducted by the school itself (Sonntag, Bassett, & Snyder, 2009). While these researchers found validity in RateMyProfessors.com, their findings were limited by the size of their study, as they focused on only one university.

Examining the medium through which the course is offered is noteworthy because it potentially eliminates several course and faculty characteristics as explanatory factors in student evaluations. The race of the instructor, for example, despite helping to explain variance in onsite classes, failed to show significant variance in online classes (Carle, 2009). Other studies explain variance in student evaluations in this area by an instructor's comfort with teaching online versus onsite (Liu, 2012). Like many of the

other characteristics in this review of the literature, there is still much to be learned in terms of student evaluations for onsite courses compared to online courses.

Class Time

The time of day a class is offered is another area of student evaluation literature. While one might hypothesize that the time the course is offered may explain variance in student evaluations, most scholars find that this is not the case (Feldman, 1978; Wachtel, 1998; Patrick, 2011). Still, there are studies such as Koushki and Kuhn (1982) that found that instructors for early morning classes received lower ratings.

Class Size

Similar to class time, class size is another course characteristic one might think would impact student evaluations. Feldman (1978) conducted a meta-analysis of student evaluations and thoroughly describes early works examining class size. He found that some studies claimed that class size had no effect on student evaluations, but that most find an inverse relationship, that is the larger the class size, the lower the student evaluation rating, and smaller the class size, the higher the student evaluation rating. Feldman goes on to explain that class size is not independent of other course variables, “for example, smaller courses at many colleges are probably more likely to be upper division courses than are larger courses” (210). The same could be said about the relationship between class size and general education or elective courses. Perhaps core courses have more students in them, which then negatively affects student evaluations.

More recent studies regarding class size and student evaluations found that the size of the class does not predict student evaluation ratings (Aleamoni, 1999; Liu, 2012; Jirovec, et al., 1998; Pepe & Wang, 2012). It should be noted, however, that Liu’s (2012)

study examined online courses, for which class size may not be as noticeable from a student's perspective.

One thing lacking in this body of literature is a consistent threshold for what constitutes a large or small class. Thresholds thus far have ranged from nine students to fifty students to differentiate large from small classes (Feldman, 1978). Scholars should consider setting a standard for the literature. This, however, can be difficult since the same sized class at two different schools might represent completely different things when compared to other classes at that school. Perhaps, then, each school should have a unique threshold, but the process for calculating the threshold when conducting research should be the same for all schools, such as average class size plus or minus a certain number of students (or even plus or minus a standard percentage from the average) for large and small courses.

Course Level

Generally, the research on course level related to student evaluations is divided into lower division, upper division, and graduate courses. Most studies of undergraduate courses find that the higher the course level, the higher the student evaluation rating (Feldman, 1978; Wachtel, 1998; Patrick, 2011). There have not been enough graduate course studies to support this claim, though course levels for graduate courses have a smaller range. Nevertheless, one study researching student evaluations in graduate courses was consistent with the majority of the undergraduate student evaluations literature in terms of faculty tenure status, faculty race, and expected grade (McPherson & Jewell, 2007).

Though there are many studies that used only upper division courses or only lower division courses, researchers often fail to explain this in their findings or limitations. This is indeed a limitation of focusing on one particular level, as it becomes difficult to compare the data to other levels.

Student Characteristics

To this point in the literature review, there have been several course and faculty characteristics discussed that could potentially sway student evaluations. What follows are student characteristics that could potentially influence student evaluations. In general, there seems to be less research on student characteristics than course and faculty characteristics, perhaps because access and reporting must consider student data privacy issues when including students.

Age

There has been some indication from the literature on course characteristics that student age affects student evaluations, most notably, through course level. Because of the findings in course level literature, one might expect to find similar results related to student age (or even year in school), that is, the younger the student the student the lower the rating. Liu (2012) confirms this assumption, noting that first year students provided the lowest ratings on student evaluations. Other scholars found correlations between student age and faculty characteristics, such as faculty age and gender (Sprinkle, 2008). Sprinkle found that older students preferred older instructors and younger students preferred younger instructors (the older/younger faculty cut off was 55 years old). Sprinkle (2008) also notes that older students preferred male instructors while younger

students preferred female instructors. However, not every study found a relationship between student age and student evaluations (Shirbagi, 2011).

Gender

Unlike what was discovered for faculty gender, most scholars agree that student gender affects student evaluations (Wachtel, 1998; Amin, 1994; Basow, et al., 2013; Bahen, McLoughlin, & Garcia, 1999; Sprinkle, 2008). While Kyei-Blankson and Nur-Awaleh (2010) found no correlation between student gender and student evaluations, the student evaluations they used evaluated the instructor's use of technology in the classroom.

Though the majority of scholars are in agreement that student gender influences student evaluations, this does not mean it happens in the same way for every study. On the contrary, some studies found that male students generally give higher ratings than female students (Basow, et al., 2013), while others found that female students rated female instructors higher and male instructors lower (Bachen, et al., 1999). This same study showed no differences in male student evaluations. It has also been reported that, while female students prefer female instructors, male students prefer male instructors (Sprinkle, 2008). The differing findings suggest that more studies are necessary to assess the impact of student gender on student evaluations, but one thing seems clear; student gender can influence student evaluations.

Anticipated or Actual Grade

Conventional wisdom tends to hold that a student's anticipated grade predicts student evaluations, with the higher the anticipated grade, the higher the student evaluation rating (McPherson & Jewell, 2007; Feldman, 1976; Caruth & Caruth, 2013).

Still, there is at least one study that shows that anticipated grade does not impact student evaluations (Centra, 2003).

The literature provides little distinction between actual grade and anticipated grade. While one is what students actually received and the other is what students expect to receive, there may be little distinction because of when the student evaluations are administered. Student evaluations are almost always collected before students receive final grades, at a time when students are likely to know what grade they will be earning. For many courses, it is easy for students to calculate their grades throughout the semester. These students not only expect to receive a certain grade, they *know* they will receive a certain grade. Still researchers have found unique ways to include actual grade in post semester studies.

Sonntag, Bassett, and Snyder (2009), for example, found that the easiness ranking on RateMyProfessors.com correlated with the actual grade the student received. It would be interesting to see how this compared to the student's anticipated grade and student evaluations rating. Brockx, Spooren, and Mortelmans (2011) also conducted a student evaluation for a course taken in the previous semester, after students had received their final grades. Their findings were in line with that of anticipated grade, the higher the grade, the higher the rating. Norvilitis and Zhang (2009) compared the actual grade for the course to an adjusted mean grade for the class. Students who scored higher than the class mean were more likely to give a better student evaluation score to the instructor.

Another important part of the literature to consider is grade inflation (Caruth & Caruth, 2013; McPherson & Jewell, 2007). Grade inflation is the belief that instructors assign higher grades than those students earned in an effort boost their student evaluation

scores, among other reasons. With so much emphasis on the use of student evaluations and given all of the biases that may exist in their completion, this remains an understandable and concerning issue for higher education institutions.

Conclusion

As outlined above, there is a significant amount of research dedicated to exploring student evaluations. Since student evaluations are so widely used in higher education for personnel decisions, it only makes sense that it be studied thoroughly to better understand to what extent faculty, course, and student characteristics might be influencing these ratings.

The literature review provided some examples of bias that existed when examining student evaluations. However, there was at least one study to the contrary (i.e., no bias) for every student, faculty, and course characteristic studied. Still, any evidence suggesting that student evaluations can be swayed is disconcerting given the significance of these evaluations.

A substantial portion of the literature reviewed was limited either in size or scope. For example, few studies examined multiple disciplines or failed to use discipline as a potential factor in student evaluations. Other studies may have had only a few hundred evaluations or only focused on one particular semester. This leads to one of the main limitations in the research reviewed, the absence of longitudinal studies. Future studies would benefit by having a multi-year, cross-discipline study so that student evaluations for a particular instructor could be measured over time and compared to other faculty during that time as well. At the same time, a study of this magnitude would benefit by including as many faculty, course, and student characteristics as accessible.

CHAPTER THREE

METHODOLOGY

This study used quantitative data analysis procedures to examine undergraduate student responses to their end-of-semester professor and course evaluations at a single college within a university located in the southeastern United States, which was called USE (University of Student Evaluations) for this study. Specifically, this study utilized data from Student Perceptions of Teaching (SPOT), an end-of-semester evaluation instrument digitally administered for each course at USE, to garner student perceptions of both their instructors and their courses. An additional database that contains information about the course itself—such as time of day and day of the week—was paired with the SPOT database to analyze the impact of contextual aspects of the course on the student evaluations. Furthermore, additional instructor variables, such as faculty gender and tenure status, were acquired through the university’s website, professional website, or social media outlet. The analysis that follows, addresses the following two research questions:

1. To what extent do student evaluations vary by academic discipline?
2. To what extent can instructor and course characteristics explain variation in student evaluations?

Research Site and Datasets

This study used student course evaluation data publicly available online from USE. This university was selected based on accessibility of its data. Since data were released into the public domain by USE, additional permission for its use was not necessary.

The following sections—Institutional Data, Student Demographics, Faculty Demographics and Course Characteristics—all used data taken from a university published document in Fall 2013. This year's document was chosen because it was published within the range of this study, Summer 2012 to Spring 2014. These data were not taken from the constructed dataset in any way and should be viewed as an informational, complimentary piece only. Specific course and faculty data from the constructed dataset used for this study will be described in the findings.

Institutional Data

USE is a Research I university. USE has over 30,000 enrolled students, with approximately 25,000 being undergraduates. Its flagship campus, the campus used for this study, is the largest and represents 80% of the university's total enrollment. The college within the university used for this study, Arts and Letters, is the third largest and represents 14% of the total enrollment at the university. The College of Arts and Letters has the most diverse course offerings of any one college at the university, which is why it was chosen for this study. The only two colleges with larger undergraduate enrollments than Arts and Letters at USE are the College of Business and the College of Science, representing 23% and 18% of the total enrollment, respectively.

Of the top ten undergraduate degree programs by enrollment, the College of Arts and Letters is only represented once in the list. This can be explained, in part, by the fact that they have the most degree programs of any of the colleges. Graduates from the College of Arts and Letters represent more academic disciplines than any other college at the university. In 2012-2013, the College of Arts and Letters awarded 1,153 undergraduate degrees, only surpassed by the College of Business, with 1,418.

Student demographics. Of the 30,000 students at USE, 94% reside within the state. 57% of the students are female and 43% are male. White students represent approximately 48% of the student population, followed by Hispanic students at 23%, Black students at 18% and Asian students at 4%. The average age is 24 for undergraduates and 33 for graduate students.

Faculty demographics. When these data were collected in the fall of 2013, there were 1,023 faculty members at USE. 549 faculty members were male and 474 female. There are 674 tenured or tenure track faculty, 415 male and 259 female. Of the 349 non-tenured and non-tenure track faculty, adjuncts, 215 were female and 134 were male.

Seventy-three percent of the tenured or tenure track faculty members at USE are White, 10.9% are Asian, 7.5% are Hispanic, and 5.1% are Black. For adjunct faculty, 80% are White, 8.3% are Black, 6.3% are Hispanic and 2.6% are Asian.

Course characteristics. At USE, there is a 25:1 student to faculty ratio. The average undergraduate class size is: 39 for a lecture course, 30 for a discussion, and 20 for a lab. For graduate courses, the average size is 13 for a lecture and 9 for a lab. The large difference in size between undergraduate and graduate class sizes is the reason why graduate courses were omitted from this study.

E-Learning is a big part of USE and there are four different types of E-Learning courses: fully online, videotaped classroom sessions, video conferencing, and mostly online. Fully online are the most popular of the four, encompassing 917 sections, 24,673 students in classes and 12,132 unduplicated student enrollment.

Datasets

The data available for student evaluations and course schedules came from two different databases and are from the six most recent semesters at the time this study began, equally representing two full years' worth of data from Summer 2012 to Spring 2014. The two databases examined were the university's course schedule database and its SPOT (Student Perception of Teaching) database. Each database is organized by semester.

From the course schedule database, the following variables were retrieved: course reference number (CRN), instructional method, the existence of additional fees, department and division, instructor, course title, days that the course was offered, time the course was offered, meeting type, term, and credits. These variables represent most of the course characteristics that were analyzed in this study.

From the SPOT database, the following variables were retrieved: number of enrolled students, response rate of student evaluations, mean of questions 19, 20, and 21, median of questions 19, 20, and 21 (calculated by the researcher), and the instructor's length of time at the institution. Questions 19, 20, and 21, which will be described in the next section, were chosen because they offered the most relevant, all-encompassing data for the study. There were some overlapping variables between the two databases, but in several cases the data were not consistent. For example, inconsistencies include course title and enrolled students. To address these inconsistencies, the course title was always taken from the course schedule database; the number of enrolled students was always taken from the SPOT database. In the case of the number of enrolled students, it seemed logical to take this data from the SPOT database because this figure more than likely

represented the number of students enrolled at the end of the semester (after adds or drops), while the course schedule database would likely represent the number of students enrolled at the beginning of the semester.

Several criteria were established for the data collected in this study. The first was with the percent of student responses for each course. Jirovec et al. (1998) maintain that a two-thirds response rate is sufficient when examining student evaluation data. As such, this study adhered to the same threshold, $\geq 66.7\%$ (the SPOT database rounds its decimals to the nearest tenth). So, the number of student responses divided by the number of students enrolled in the class needed to be greater than or equal to .667. As described in the literature, this two-thirds threshold was established because courses with lower response rates may be more prone to biased responses, and, as a result, are less likely to give an adequate representation of the students' perceptions of the instructor or course. There was one exception to the two-thirds threshold in this study, and it was if the number of student responses (x) was ≥ 50 , then the class was included if the overall response rate was 50% or more.

Second, courses needed to have a minimum number of responses to be eligible. The minimum number of responses was set to ten as suggested by Marsh (1982). This effectively eliminated smaller courses and independent studies, while still provided enough responses to have relevance. Additionally, there was no cap on the size of the class, since previous scholars have explored the impact of class size on student evaluations (McPherson & Jewell, 2007; Zhao & Gallant, 2012; Pepe & Wang, 2012; Wachtel, 1998; Jirovec et al., 1998; Aleamoni, 1999; Feldman, 1978).

Third, all data collected were for undergraduate courses at one college (Arts and Letters) on one campus at the university. The campus was chosen because it is the university's largest and most diverse in terms of majors. Similarly, the College of Arts and Letters was chosen because it represents the greatest diversity of course offerings from any one college at the university.

Finally, courses that had "staff" as the instructor for the course were omitted from the analysis because it was impossible to know exactly which faculty member actually taught the course. These four criteria were used to identify eligible courses, which were then linked to the characteristics of both the course and the faculty member that taught the course, and then used to help explain variation in the student evaluations.

Student Perceptions of Teaching (SPOT) Instrument

This study utilized data from the Student Perceptions of Teaching (SPOT) database composed of student evaluations collected in an online instrument of the same name. The instrument solicits eighteen different responses from students, fifteen of which use a four-point Likert scale ranging from "Strongly Agree" to "Completely Disagree." The final three questions utilized a five-point Likert scale with different options for each question (See Appendix A for a sample evaluation). The SPOT instrument is administered online for every class at the end of the semester. This was a major benefit for choosing this research site, since many universities, or even colleges within a university, do not use a standardized evaluation instrument like USE does.

SPOT responses are anonymous and do not include any student data, with the exception of the number of students in the class (y), number of students who completed the survey (x) and percent of students who participated in the survey (x/y; n). As

previously mentioned, to be eligible for this study, courses had to fit within the following criteria: the number of student responses had to be at least 10 and the response rate had to be $\geq 66.7\%$, or, the number of student responses had to be at least 50 and the response rate had to be $\geq 50\%$. Large classes typically had lower response rates, so this exception was made to include large classes in the study without ultimately impacting the integrity of the study since, with this lower threshold, came a larger number of responses.

Of the eighteen Likert-scaled questions shown on the SPOT database for each course, only three are used for the current study. These three, questions 19-21, were chosen based on their overarching relevance to the research questions for this study. The three items used for this study are as follows (see Appendix A for a visual representation of the survey):

19. How much do you think that you have learned in this course?
20. Rate the quality of instruction as it contributed to your learning in the course.
21. What is your rating of this instructor compared to other instructors you have had?

In the interest of parsimony, these questions will be shortened when they are referenced in the coming chapters. Question 19 will be referred to as student learning. Question 20 will be referred to as instructional quality. Question 21 will be referred to as instructor comparison.

These questions were scored from one to five on a Likert scale, with one being the score associated with the best option and five being associated with the worst. For each of these items, the mean score was noted in the compiled dataset for this study. In addition, the median was also calculated, since it was expected to serve as a better

measure for these student evaluations. The median is likely to be a better measure than the mean because it effectively eliminates the outliers, such as students who rated an instructor poorly because they were upset with the instructor. Similarly, it also prevents students who usually just give perfect scores to all their instructors from inflating the instructor's average score. The difference between the mean and median will be noted in the findings. These mean and median scores served the focal point for comparing course and faculty characteristics in the dataset.

Course Schedule

The second database used in examining the research questions for this study was the course schedule database. Unlike SPOT, there was no evaluation attached to this database, it was merely a compilation of every course offered from Summer 2007 to Fall 2014. This database served as the first step in finding eligible classes for this study. Variables included in the course schedule dataset were: term, campus, college, department, level, and instructional method, all of which were relevant variables that were included for each course. Campus, college, and level were the only three factors that remained constant for this study. The term ranged from Summer 2012 to Spring 2014. This study only examined departments within the College of Arts and Letters as it was picked based on the diversity of its course offerings.

Department was another variable included in this dataset and it included eleven of the possible seventeen departments. Six departments were omitted as a result of insufficient course offerings and eligible data for this study. The departments used for this study were: Anthropology, Art, Communications and Multimedia, English, History, Languages Linguistics Comparative Literature, Music, Philosophy, Political Science,

Sociology, and Theatre and Dance. In terms of instructional method, the university used in this study had several modes of course offerings: fully online, in-person, mostly/mixed online, and video streaming. Of these four instructional methods, only two were found in the eligible course data, in-person and mostly/mixed online. Additional variables included in the constructed database were course number as a proxy for level (upper vs. lower division), days of the week the class met, and time of day the class met.

Though actual enrollment was available in this database, the enrollment numbers from the SPOT database were used since those numbers more accurately reflect the number of students enrolled in the course at the end of the semester. This was typically a lower number than that of the course schedule database, indicating that some students dropped the course at some point during the semester.

Additional Resources

While the course data was readily available, faculty data were not so easily attained. Still, these faculty data served an important purpose in determining if faculty characteristics had an impact on student evaluations. These data included: gender, length of service at the school, and tenure status.

Gender was identified by pronouns used on the institution's "faculty profile" page or other pages published by the institution. If gender was not determined this way, social media such as LinkedIn, Facebook, or other professional pages were used. If there were no pronouns used, the faculty were omitted from the study.

Faculty for whom tenure status could not be determined were also omitted from this study. These data were gathered from the university's website, other professional websites, and social media. There were only two distinctions for this faculty

characteristic, tenured/tenure-track and non-tenure track. Tenured and tenure-track status was identified using the institution's published rankings. Assistant professor, associate professor, and professor were the three levels of tenured/tenure-track at the USE. All other titles in this study were considered non-tenured/tenure track and included, but were not limited to: instructor, visiting instructor, and adjunct faculty.

Length of time at the university was a faculty characteristic included in the constructed dataset. Specifically, had the instructor been at the university for at least five years at the time that their course was listed in the dataset? Adding this variable was thought to shed light on the relationship between experience at a single university and ratings on student evaluations. All the data used to determine length of time at the university were taken from either the Spring 2009 semester, or Fall 2008 semester, if the instructors were not listed in the Spring. Two semesters were used in case an instructor was not teaching during the Spring semester. The Spring 2009 semester was chosen because it represented five years prior to the start of the data collection for this study, Spring 2014.

CHAPTER FOUR

ANALYSIS

The purpose of this study was to contribute to the existing literature on student evaluations by examining the impact that course and instructor characteristics had on student evaluations. This study employed a sample drawn from two publically available, complimentary databases from a single university that allowed for the collection of most of the variables used in the study. This chapter begins with an extended discussion of the sample used for this study, followed by the results of the analysis ordered by research question. In addressing the research questions, numerous tables will be presented and their content briefly discussed.

Sample Characteristics

The completed database consisted of 1,812 unique course observations and 56 variables, though many of the variables overlapped and were added later as dummy variables for use in the analysis. Of the 1,812 courses, approximately 60% were taught by females (n=1,090) while about 40% were taught by male instructors (n=717); unfortunately, the gender of five instructors could not be determined. Taken together, there were 402 different faculty members that taught 454 different courses included in the sample. Of the 402 faulty, approximately 58% were female (n=233) and 42% were male (n=169). These numbers differ from the institutional averages of 54% male and 46% female faculty. However, since this study focuses on one college within the university, the sample in the constructed database may still be representative of the studied college's faculty, but this could not be determined. There were 11 departments included in this sample and the number of courses that were included from each department during the

studied time period can be seen in Table 1. This table shows that English and Language make up more than 50% of the sample, helping to confirm the need to analyze the sample by department to avoid oversaturation by these two departments in the sample.

Table 1

Total Courses Per Department

Department	Frequency	Percent
Anthropology	72	4.0
Art	153	8.4
Communications and Multimedia	164	9.1
English	542	29.9
History	84	4.6
Language Linguistics Comp Lit	372	20.5
Music	104	5.7
Philosophy	74	4.1
Political Science	85	4.7
Sociology	104	5.7
Theatre and Dance	58	3.2
Total	1812	100.0

Research Question 1

The first research question was: to what extent do student evaluations vary by academic discipline? To answer this question, the use of descriptive statistics was required. Table 2 represents the descriptive statistics of the six student evaluation variables for every course observation (n=1,812). These six student evaluation variables were the mean and median for the SPOT research questions on student learning, instructional quality and instructor comparison.

Table 2

Descriptive Statistics for All Observations (All Departments)

SPOT Questions	Mean	Median	Range	Standard Deviation
Question 19: How much do you think you have learned in this course?	2.09	2.00	3.19	.43
Question 20: Rate the quality of instruction as it contributed to your learning in the course	1.82	1.50	3.38	.51
Question 21: What is your rating of this instructor compared to other instructors you have had?	1.97	2.00	3.34	.52

An examination of Table 2 reveals relatively low means and medians, indicative of good student evaluation scores. However, the two measures of dispersion associated with the questions – the range and standard deviation – reveal significant variation that will be explored later in the analysis surrounding the second research question. However, to better understand this table and those that follow, recall that students could rate the SPOT evaluation questions from 1-5, with 1 being the best score and 5 being the worst. Though each question is associated with a unique set of descriptors, these ratings nevertheless mean the same thing. For example, the three questions all treat a score of 1 or 2 as being above average, a score of 3 being average and a score of 4 or 5 as being below average. The lowest (best) score was for the instructional quality question (#20), followed by the instructor comparison question (#21). Although still a good score, the highest (worst) of the three questions was associated with student learning (#19). As an example, a single course evaluation for each of these questions as well as the associated descriptors are presented in Tables 3-5.

Table 3

SPOT Database Question 19 Example

	An Exceptional amount (1)	More than usual (2)	About as much as usual (3)	Less than usual (4)	Almost nothing (5)	No response	Mean
19. How much do you think that you have learned in this course?	46.2	30.8	15.4	7.7	0.0	0.0	1.85

Table 4

SPOT Database Question 20 Example

	Excellent (1)	Very Good (2)	Good (3)	Fair (4)	Poor (5)	No response	Mean
20. Rate the quality of instruction as it contributed to your learning in the course.	46.2	15.4	23.1	7.7	0.0	7.7	1.92

Table 5

SPOT Database Question 21 Example

	One of the most effective (1)	More effective than others (2)	About as effective as others (3)	Less effective than others (4)	One of the least effective (5)	No response	Mean
21. What is you rating of this instructor compared to other instructors you have had?	46.2	30.8	23.1	0.0	0.0	0.0	1.77

Taken together, the sample student evaluation shown in Tables 3-5 display above average ratings, with all three means below the numerical score of two, which was consistent with the overall results. Interestingly, most of the courses used in this study had values of zero in the no response column. However, for those for which students selected no response, they were typically low numbers that indicated there only one or two students in the course had declined to answer a particular question. There were no evaluations that were missing enough data to warrant exclusion from the study.

An important part of addressing the first research question is separating the data by department. Tables 6 and 7 show the descriptive statistics for each department from the same six variables in the dataset pertaining to the three SPOT questions. Table 6 presents the means and Table 7 the medians for the three questions. Although not calculated as part of the SPOT database, the median was also calculated because, as discussed in the previous chapter, the median may have several advantages over the mean in databases of this type. An examination of these tables reveal that Political Science had

the lowest (best) average scores, while Philosophy had the highest (worst) scores. Still, no departments had an overall mean or median more than 2.50, suggesting that these departments in the College of Arts and Letters are above average. Future studies would benefit by including departments in other schools on campus to determine if the results are comparable.

Table 6

Means and Standard Deviations by Department for the Three SPOT Questions

Department	Student Learning		Instructional Quality		Instructor Comparison	
	Mean	SD	Mean	SD	Mean	SD
Anthropology	2.19	.40	2.02	.54	2.16	.54
Art	1.95	.44	1.80	.51	1.97	.51
Communications	2.19	.40	1.96	.52	2.11	.51
English	2.22	.42	1.88	.52	2.02	.52
History	1.95	.39	1.72	.46	1.88	.51
Language	1.99	.36	1.75	.45	1.89	.46
Music	2.01	.44	1.70	.43	1.83	.47
Philosophy	2.28	.56	2.05	.64	2.22	.64
Political Science	1.80	.39	1.64	.39	1.69	.41
Sociology	2.05	.41	1.77	.49	1.90	.53
Theatre	1.95	.45	1.70	.50	1.90	.52

Table 7

Medians and Standard Deviations by Department for the Three SPOT Questions

Department	Student Learning		Instructional Quality		Instructor Comparison	
	Median	SD	Median	SD	Median	SD
Anthropology	2.00	.54	2.00	.78	2.00	.74
Art	2.00	.64	1.50	.66	2.00	.70
Communications	2.00	.55	2.00	.68	2.00	.72
English	2.00	.59	2.00	.70	2.00	.71
History	2.00	.56	1.00	.62	1.50	.68
Language	2.00	.53	1.00	.62	2.00	.67
Music	2.00	.68	1.00	.63	1.00	.71
Philosophy	2.00	.78	2.00	.87	2.00	.88
Political Science	2.00	.56	1.00	.56	1.00	.59
Sociology	2.00	.60	1.00	.64	2.00	.70
Theatre	2.00	.65	1.00	.72	2.00	.71

Research question one also requires the inferential use of analysis of variance techniques (ANOVA) and the non-parametric Kruskal-Wallis test to statistically compare the means and medians associated with the three SPOT questions across departments. Examination of Table 8 and Table 9 reveals that the means and medians for these questions are not statistically the same at the $p \leq .05$ level across departments, indicating that the analysis for the second research question should be conducted solely at the departmental level as well.

Table 8

One-Way ANOVA for the Three SPOT Questions Based on Department

		Sum of Squares	Df	Mean Square	F	Sig.
Mean Student Learning by Department	Between Groups (Combined)	31.58	10	3.16	18.45	.000
	Within Groups	308.20	1801	.17		
	Total	339.78	1811			
Mean Inst. Quality by Department	Between Groups (Combined)	19.49	10	1.95	7.89	.000
	Within Groups	444.79	1801	.25		
	Total	464.28	1811			
Mean Inst. Comp. by Department	Between Groups (Combined)	24.77	10	2.48	9.64	.000
	Within Groups	462.70	1801	.26		
	Total	487.47	1811			

$p \leq .05$

Table 9

Kruskal-Wallis Test of Medians for the Three SPOT Questions by Department

	Student Learning	Instructional Quality	Instructor Comparison
Chi-Square	143.16	67.73	94.73
Df	10	10	10
Asymp. Sig.	.000	.000	.000

$p \leq .05$

Research Question 2

The second research question in this study was: to what extent can instructor and course characteristics explain variation in student evaluations? This question lends itself to using multiple regression analysis. The means and medians for the SPOT questions will be used as the dependent variables in the analysis and the collected course and faculty characteristics as the independent variables.

Although a natural starting place for this analysis might be to run a regression with these departments together, unfortunately, the selection criteria used to draw the sample meant that the courses selected were not naturally representative of anything within the College of Arts and Sciences. When combined with the results from the ANOVA and the Kruskal-Wallis tests, research question one that showed that the means and medians for each department were significantly different, the second research question will take the perspective of the individual department.

The first part of addressing this research question focuses on faculty data used in the sample. While there were numerous faculty data points that would have been relevant to this study, such as, age, race and ethnicity, unfortunately these variables were inaccessible. Still, three faculty variables could be determined and were included in the sample: gender, tenure status, and time at the university. While the amount of time an instructor had been at the university could be determined using the SPOT database, the other variables could not. Instead, for gender and tenure status, the university's website, the instructor's professional website and even social media was used to determine these variables. As mentioned previously, the instructor's gender was never assumed. If there

were no pronouns used to identify the instructor's gender, then the instructors were omitted from the study.

Sixty-six different linear regressions were conducted (11 departments x 6 dependent variables) and the results show that all faculty characteristics were significant for each question variable in at least one department ($p \leq .05$). Tables 9 and 10 illustrate the faculty characteristics regression based on the means and medians of the three SPOT questions. The estimated coefficients displayed represent how the variable positively or negatively impacts the dependent variable (SPOT questions). It is important to remember again, that, in the case of these student evaluations, the lower the mean or median, the better rating it represents. This means that the negative estimated coefficients displayed for variables actually impact student evaluations for the better.

The independent variables were dummy variables. In the sample, for gender, a value of one indicated the course was taught by a female and a zero indicated it was taught by a male. Thus, in terms of the regression in Tables 10 and 11, the figures presented for gender only consider female instructors. If the figure is positive for a department, then that means female instructors in that department received lower ratings than male instructors. If the figure is negative, then that means female instructors in that department received better ratings than male instructors. The other independent variables, an instructor teaching at the university for at least five years and an instructor's tenure status, were also dummy variables. A one in the sample indicates that an instructor has taught at the university for at least five years, while a zero in the sample indicates that the instructor has not taught at the university for at least five years. Similarly, a one indicates that the instructor has tenure or is a tenure-track faculty

member, while a zero indicates that the instructor is not a tenure-track faculty member.

In examining the figures in the table for the five years and tenure variables, a positive figure indicates that an instructor with tenure or tenure-track or five years teaching at the same university had an adverse effect on student evaluations, while a negative figure indicates that an instructor with tenure or five years teaching experience at the university received better ratings than instructors without tenure or less than five years of experience teaching at the university.

Table 10

Faculty Characteristics Regression Based on Means of the Three SPOT Questions by Department

Dept.	Student Learning			Instructional Quality			Instructor Comparison		
	Gender	5 Years	Tenure	Gender	5 Years	Tenure	Gender	5 Years	Tenure
Anthropology		.31*		.44*	.39*		.52**	.47*	
Art									
Communications				-.24**			-.23**		
English	.13***	.12***	-.19***	.23***	.15***		.18***	.17***	
History		-.45*	.40*			.44*			
Language	.13**						.13*		
Music				.26**			.29**		
Philosophy									
Political Science				-.28*					
Sociology		-.33***		-.21*	-.36**		-.30**	-.37**	
Theatre & Dance			-.33*					.45*	-.48**

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 11

Faculty Characteristics Regression Based on Medians of the Three SPOT Questions by Department

Dept.	Student Learning			Instructional Quality			Instructor Comparison		
	Gender	5 Years	Tenure	Gender	5 Years	Tenure	Gender	5 Years	Tenure
Anthropology		.40*			.24*		.69*	.65*	
Art									
Communications				-.41***			-.33**		
English	.15**	.19***	-.27***	.27***	.13*	-.15*	.18**	.24***	
History	.30*	-.86**	.71**						
Language	.17*								
Music				.35**			.44**		
Philosophy									
Political Science				-.41*					
Sociology	-.28*	-.60***					-.44**	-.38*	
Theatre & Dance								.54*	-.56*

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Taken together, gender had the most significant estimated coefficients.

Depending on the department, however, this value could be positive or negative. Because this variable was defined as one for women and zero for men, negative coefficients were associated with higher scores for women while positive coefficients were associated with higher scores for men. Notably, the tenure/tenure-track variable had the fewest significant estimated coefficients of the three faculty characteristics used in the models.

Before discussing the models themselves, it is important to note that the correlations among these faculty variables were modest, with the largest being .38 between tenure or tenure-track and five years as shown in Table 12.

Table 12

Faculty Characteristics Correlations

	Gender	5+ Years	Tenure
Gender - Pearson Correlation	1	-.098**	-.209**
Sig. (2-tailed)		.000	.000
N	1807	1807	1796
5+ Years - Pearson Correlation	-.098**	1	.377**
Sig. (2-tailed)	.000		.000
N	1807	1812	1801
Tenure - Pearson Correlation	-.209**	.377**	1
Sig. (2-tailed)	.000	.000	
N	1796	1801	1801

** $p \leq .01$

While it makes sense that these two variables (tenure status and length of time at the university) were positively correlated, the fact that the correlation is so low suggests that there are a good number of non-tenured faculty who have been at the university for five or more years. In general, correlations are considered small up to .3, medium up to

.5 and large up to 1.0. None of the correlations between faculty characteristics were large.

For the course characteristics correlations, on the other hand, there were several significant, large values, but they were expected, as they were the days of the week. For example, Monday, Wednesday, and Friday, were significantly correlated, as were Tuesday and Thursday. This makes sense, given the fact that many classes are typically offered on Monday, Wednesday, and Friday, or on Tuesday and Thursday.

Tables 13 through 18 illustrate the course characteristics regressions based on the means and medians of the three SPOT questions. The estimated coefficients represent how the variables positively or negatively impact the model's dependent variable (the means and medians for the three SPOT questions). Sixty-six different linear regressions were done (11 departments x 6 dependent variables) and the results show that many, but not all the course characteristics were significant for each question variable ($p \leq .05$). Instructional method, lecture, STU (studio), and PRF (performance) showed no significance for any question or department.

Table 13

Estimated Coefficients and Levels of Significance for Course Characteristics Regression Based on the Mean from the Student Learning Question

Dept	Fee	Division	Mon	Tues	Wed	Thurs	Friday	Early AM	After noon	Eve	Fall	Spring	GenEd
Anthro													
Art	-.37**												
Comm													
English				-.37*				.11*		-.18*			
History		-1.43***											-1.07*
Lang	.20*	.22**									.18**	.17*	.65*
Music										.31*			.62**
Philo					-1.47*	-1.38*	-1.66***		-.34*		.51**		
PoliSci							-.42**			-.40*			
Soci			.81**										-.21*
Theatre													

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 14

Estimated Coefficients and Levels of Significance for Course Characteristics Regression Based on the Mean from the Instructional Quality Question

Dept	Fee	Division	Mon	Tues	Thurs	Friday	Afternoon	Eve	Discussion	Fall	Spring	GenEd
Anthro		1.03*		-.60*						.57*		1.16**
Art	-.48**											
Comm												.56*
English					-.51*							
History		-1.45**	-.61*						-1.14*			-1.14*
Lang										.24**	.26**	
Music												.46*
Philo						-2.01***	-.45**			.50*		
PoliSci								-.37*				
Soci			.80*									
Theatre												

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 15

Estimated Coefficients and Levels of Significance for Course Characteristics Regression Based on the Mean from the Instructor Comparison Question

Dept	Fee	Division	Mon	Tues	Wed	Thurs	Friday	Evening	Lab	Fall	Spring	GenEd
Anthro		.84*		-.71*					-.67*	.48*		1.01*
Art	-.52***											
Comm												.58**
English				-.47*	-.81*	-.69**		-.26*				
History		-1.58**	-.66*							-.45*	-.52*	-1.38*
Lang										.22*		
Music												
Philo							-1.79**					
PoliSci							-.41**	-.42*				
Soci			.72*									-.28*
Theatre												

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 16

Estimated Coefficients and Levels of Significance for Course Characteristics Regression Based on the Median from the Student Learning Question

Dept	Fee	Division	Mon	Tues	Wed	Thurs	Fri	Early AM	After noon	Eve	Spec ial	FA	GenEd
Anthro													
Art	-.47*												
Comm													
English				-.55*		-.63*		.19*					
History		-1.73**					.54*						-1.39*
Lang	.32*	.33**									-.39*		.81*
Music	.45*									.43*			1.03**
Philo					-2.50**	-2.50**	-2.39***		-.38*			.64**	
PoliSci							-.68**			-.73**			
Soci			1.15**										-.38**
Theatre													

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 17

Estimated Coefficients and Levels of Significance for Course Characteristics Regression Based on the Median from the Instructional Quality Question

Dept	Fee	Division	Mon	Thurs	Fri	Afternoon	Fall	Spring	GenEd
Anthro		1.32*							1.53*
Art	-.53*								
Comm									.84**
English			-.46*	-.78*					
History		-2.02**							-1.70*
Lang							.36**	.42***	
Music								.64*	.94**
Philo					-2.57**	-.54*	.72*		
PoliSci					-.44*				
Soci									
Theatre									

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 18

Estimated Coefficients and Levels of Significance for Course Characteristics Regression Based on the Median from the Instructor Comparison Question

Dept	Fee	Division	Tues	Wed	Thurs	Fri	Eve	Lec	Lab	Special	FA	Spring	GE
Anthro		1.20*	-1.11**						-1.04*				1.42*
Art	-.52*									-.95*			
Comm													.75*
English			-.96***	-1.16**	-.76*		-.38**						
History		-2.07**									-.70*	-.73*	-1.91*
Lang													
Music													.77*
Philo						-2.05*					.74*		
PoliSci						-.58**	-.74**						
Soci													-.38*
Theatre								.87**					

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

In constructing Tables 13-18, four variables were omitted to avoid multicollinearity: weekend, late morning, LLB (Bachelor of Law Course), and Summer. As such, the estimated coefficients for the variables included in the analysis reflect their differences with the relevant omitted category. This means, for example, that a positive value for a course in the fall indicates a higher score (worse) when compared to the summer. General education courses showed up most frequently in Tables 13-18 and had both positive and negative estimated coefficients depending on the department. While division didn't appear in the tables too frequently, when it did it often had large coefficients, and, like general education, had both positive and negative values, depending on the department.

In Tables 19-29, the results are shown for these regressions for all course and faculty characteristics stepwise by department. The three studied SPOT questions on Student Learning, Instructional Quality, and Instructor Comparison will be abbreviated as SL, IQ, and IC, accordingly, to make the tables more concise. Of the 27-different course and faculty characteristics, nine variables failed to display significant coefficients in any of the questions/departments: instructional method, Wednesday, weekend, afternoon, lecture, discussion, lab, LLB, and PRF.

Table 19

Regression for All Course and Faculty Characteristics Stepwise for Anthropology

	Mean SL	Mean IQ	Mean IC	Median SL	Median IQ	Median IC
Fall	.24*		.32*	.28*		
Additional Fee			-.38*		-.54*	-.71**
Gender						.56*

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 20

Regression for All Course and Faculty Characteristics Stepwise for Art

	Mean SL	Mean IQ	Mean IC	Median SL	Median IQ	Median IC
Additional Fee	-.37***	-.42***	-.43***	-.42***	-.46***	-.46***
Division			-.26**		-.28*	-.42***
5 Years					.28**	

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 21

Regression for All Course and Faculty Characteristics Stepwise for Communications

	Mean SL	Mean IQ	Mean IC	Median SL	Median IQ	Median IC
GenEd	.32***	.50***	.54***		.47**	.43*
Gender					-.32**	-.25*

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 22

Regression for All Course and Faculty Characteristics Stepwise for English

	Mean SL	Mean IQ	Mean IC	Median SL	Median IQ	Median IC
GenEd	.26***	.12*		.35***	.20**	.24***
5 Years	.12***	.15***	.17***	.18***	.13*	.24***
Gender	.11**	.22***	.17***	.11*	.26***	.16*
SP	-.22*					
Division			-.10*			
EarlyAM				.14*		
Tuesday						-.13*

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 23

Regression for All Course and Faculty Characteristics Stepwise for History

	Mean SL	Mean IQ	Mean IC	Median SL	Median IQ	Median IC
Division	-1.34***	-1.25**	-1.48**	-.44***	-.67***	-1.99***
GenEd	-.89**	-.81*	-1.04*			-1.43*
Tenure	.20*	.41***	.35**		.52***	.54***

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 24

Regression for All Course and Faculty Characteristics Stepwise for Languages

	Mean SL	Mean IQ	Mean IC	Median SL	Median IQ	Median IC
Friday	.11**			.18**		
Summer	-.16**	-.23**			-.28**	
Gender	.12*		.13*			
EarlyAM		.13*				
Fall			.10*			
SP				-.37*		
Evening					-.21*	-.25*

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 25

Regression for All Course and Faculty Characteristics Stepwise for Music

	Mean SL	Mean IQ	Mean IC	Median SL	Median IQ	Median IC
LateAM	-.37***	-.28***	-.22*	-.36**		-.29*
Gender	.32***	.41***	.41***		.49***	.63***
Division	-.25***		-.22**			-.34**
Additional		-.32***				
Fee						
STU			-.32***		-.36**	-.42**
Thursday			.19*	.28*		
GenEd				.88***	.60*	
EarlyAM					.31*	

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 26

Regression for All Course and Faculty Characteristics Stepwise for Philosophy

	Mean SL	Mean IQ	Mean IC	Median SL	Median IQ	Median IC
GenEd	1.11***			1.13***		
Friday	-.36*	-.74***	-.58**		-1.11***	-.82**
Division		-1.05***	-1.05***		-1.61***	-1.55***
LateAM		.34*			.45*	
Summer					-.64*	-.60*

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 27

Regression for All Course and Faculty Characteristics Stepwise for Political Science

	Mean SL	Mean IQ	Mean IC	Median SL	Median IQ	Median IC
Friday	-.49***	-.35**	-.42***	-.57***	-.50**	-.67***
Evening	-.36**	-.42***	-.43***	-.67***	-.62***	-.76***
Gender	-.24**	-.29***	-.35***		-.42***	-.58***
Division	-.21*					
Summer					-.35*	
Spring						-.28**

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 28

Regression for All Course and Faculty Characteristics Stepwise for Sociology

	Mean SL	Mean IQ	Mean IC	Median SL	Median IQ	Median IC
Monday	.46***	.50***	.64***	.66***	.60***	.65***
Fall			.27**		.24*	.31*
Division			.19*			
5 Years				-.40***		
Tenure				.37**		

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Table 29

Regression for All Course and Faculty Characteristics Stepwise for Theatre and Dance

	Mean SL	Mean IQ	Mean IC	Median SL	Median IQ	Median IC
Fall	-.29*	-.27*	-.37**	-.51**		-.52**
STU		-.25*				
Spring					.42*	

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Some departments, such as History, had large, significant coefficients, with one estimated coefficient being -1.99 as noted in Table 23. The larger the coefficient, the greater the variation the characteristic helps explain. Gender, division, and general education were the most common significant variables across the 11 departments. In Tables 19-29, gender was significant in 21 of the possible 66 regressions, while general education and division showed up as significant in 18 and 19 of the 66 regressions, respectively. Seven of the statistically significant gender coefficients were negative and 14 were positive, suggesting that male instructors tended to score better than female instructors. For general education, 14 of the statistically significant coefficients were positive compared to four negative. This suggests that general education courses typically receive lower ratings than non-general education courses. Finally, for division, 18 of the 19 statistically significant coefficients were negative, suggesting that upper division courses typically receive better ratings than lower division courses.

In an effort to make these effects easier to see, Table 30 was created as a way of summarizing Tables 19-29 to show all departments and their corresponding significant estimated coefficients. Next to each course and faculty variable listed under the significant coefficients column is either a "+" or "-" sign. This indicates whether the

estimated coefficients for the described coefficients listed in Tables 19-29 were positive or negative.

Table 30

Summarized Regression for All Course and Faculty Characteristics for All Departments

Department	Significant Coefficients
Anthropology	Fall (+), Additional Fee (-), Gender (+)
Art	Additional Fee (-), Division (-), 5 Years (+)
Communications	GenEd (+), Gender (-)
English	Gen Ed (+), 5 Years (+), Gender (+), SP (-), Division (-), EarlyAM (+), Tuesday (-)
History	Division (-), GenEd (-), Tenure (+)
Language	Friday (+), Summer (-), Gender (+), Early AM (+), Fall (+), SP (-), Evening (-)
Music	LateAM (-), Gender (+), Division (-), Additional Fee (-), STU (-), Thursday (+), GenEd (+), EarlyAM (+)
Philosophy	GenEd (+), Friday (-), Division (-), LateAM (+), Summer (-)
Political Science	Friday (-), Evening (-), Gender (-), Division (-), Summer (-), Spring (-)
Sociology	Monday (+), Fall (+), Division (+), 5 Years (-), Tenure (+)
Theatre and Dance	Fall (-), STU (-), Spring (+)

As shown in Table 30, the variable additional fee always displayed a negative coefficient, meaning, despite paying more for the class, for every applicable department, students gave these courses better scores. In terms of the time the course was offered and how that impacted student evaluations, Table 30 shows a preference for evening courses, with both mentions being represented by a negative significant coefficient. Early

morning courses, on the other hand, appeared only as a positive significant coefficient, which resulted in poorer marks on student evaluations.

Another interesting finding was that tenure status, when statistically significant, was always represented by a positive coefficient for tenured and tenure track faculty. Students rated instructors who were tenured or tenure-track lower than those not on tenure track. Time of year was also an interesting. Summer courses only appeared as a negative significant coefficient, Spring appeared equally as both a positive and negative significant coefficient and Fall was a positive significant coefficient, in three out of the four departments where it showed up.

There were seven different “meet types” on the course schedule database. The only two that had statistically significant coefficients were STU (Studio) and SP (Special Topics) and they were always negative when they were significant meaning students liked these courses.

Overall, Communications had the fewest unique significant coefficients, two, while Music accounted for the most significant coefficients, eight. Finally, no departments had both positive and negative coefficients for the same variable, which is expected, but nevertheless confirms the methodology used for this study.

CHAPTER FIVE

DISCUSSION

This chapter will begin with a brief description of the purpose of the study and methodology used for this study. After reviewing the methodology, the findings will be elaborated on and discussed with regard to how they fit into the literature. This section will be followed a description of the implications of the findings, with particular attention given to policy implications. The chapter will end with a section on the limitations of this study and finally a discussion of the implications for future research.

Purpose of the Study

In consideration of the literature reviewed, this study was designed to examine two things: the extent to which student evaluations vary both within and across disciplines, and the extent to which instructor and course characteristics are associated with variation in student ratings. These points of interest were addressed in the following two research questions:

1. To what extent do student evaluations vary by academic discipline?
2. To what extent can instructor and course characteristics explain variation in student evaluations?

Methodology Review

Before discussing the two research questions and their findings, the methodology used for this study will be reviewed. This study utilized the SPOT database and course and faculty data from one college, on one campus, at one university in the southeastern United States (USE). All collected data were public, which was the main consideration in choosing USE. The sample represents two years of data, six total semesters, Summer

2012 to Spring 2014. In total this study collected course and faculty data for 1,812 individual courses ($n=1,812$). There was a total of 11 different departments included from the college selected, the College of Arts and Letters, and there were 27 different course and faculty variables collected for each observation.

As mentioned in previous chapters, not all courses were eligible to be included in this study. In addition to already being within the studied school/campus/university, the number of student evaluations collected for a particular course needed to be greater than or equal to two-thirds of the total number of enrolled students, as recommended by Jirovec et al. (1998). In addition, there needed to be at least ten completed evaluations for a course to be included, which Marsh (1982) suggested was a good minimum. While the minimum number of students was set at ten, there was no maximum included since researchers have suggested class size can impact student evaluations (McPherson & Jewell, 2007; Zhao & Gallant, 2012; Pepe & Wang, 2012; Wachtel, 1998; Jirovec et al., 1998; Feldman, 1978). Furthermore, many of the large courses in this study failed to meet the two-thirds response rate threshold. As such, an exception was made to ensure some larger classes were included in this study. If a course had 50 or more completed student evaluations, then the response rate could be 50% or more instead of 66.7% or more. Approximately half of the courses within this school/campus/university were excluded from this sample because they failed to meet one or more of the above-mentioned criteria.

The first research question, which explored how student evaluations varied by discipline, used both descriptive and inferential statistics. The mean, median, range, and standard deviation were collected for all courses and were examined based on

department. Political Science had the lowest (best) scores and Philosophy had the highest (worst) scores. Both analysis of variance (ANOVA) and non-parametric tests were then used to determine if the means and medians were the same by department. The results showed that the means and medians were not the same and prompted the decision to explore the second research question by department as well.

The second research question explored course and faculty characteristics from the sample that were correlated with the student evaluations. This question, as mentioned, was examined at the department level as well and required the use of multiple regression analysis. There were six different dependent variables used (the mean/median for the three different SPOT questions), and 27 different independent variables that described course and faculty characteristics. Each regression was run separately by department. While some of the course and faculty characteristics showed no significant impact on student evaluations, the majority did at least once. These findings will be discussed thoroughly in the next section along with how they relate to the literature.

Discussion of Findings

This section will examine the findings presented in chapter four and, where applicable, how they relate to the rest of the literature. The findings discussed in this section will be organized by research question.

Question 1

This section is devoted to a thorough discussion of the findings for the first research question: To what extent do student evaluations vary by academic discipline?

Two departments, English and Language Linguistics Comp Lit, comprised more than 50% of the sample, 29.9% and 20.5% respectively, as shown in Table 1. This is

largely attributed to the general education courses required in these departments. Approximately 72% of the English courses in this study (392) fulfilled general education requirements, while Language Linguistics Comp Lit likely had as many general education courses since there is a foreign language requirement for most, if not all, majors at the University. However, since students could have potentially taken an additional language course as an elective or as a requirement for their language major, the number of general education courses in the sample from this department is skewed. Philosophy represented the highest percent of general education courses with approximately three-quarters of the courses for Philosophy in the sample fulfilling a general education requirement.

Overall, the courses/instructors included in this study were above average according to the descriptive statistics results in Table 2. This finding is consistent with the literature that found that humanities, art, and social sciences courses typically received better ratings on student evaluations than math and science courses (Feldman, 1978; Cashin, 1995; Centra, 1993). Since this study only examined the College of Arts and Letters, Feldman's (1978) work may help explain why the sample for this study produced above average results.

For all observations, the SPOT question on student learning had a mean score of 2.09 and a median score of 2.00. The SPOT question on instructional quality had a mean score of 1.82 and a median of 1.50. Finally, the SPOT question on instructor comparison had a mean score of 1.97 and a median score of 2.00. As noted, the lower the score, the better, and a score of approximately 2 or lower means students, in general, thought their instructor/course was above average as evidenced by the available responses in Tables 3,

4, and 5. One interesting finding, however, is the range for these questions. The three studied SPOT questions had ranges of 3.19 (student learning), 3.38 (instructional quality), and 3.34 (instructor comparison). This shows that not all students had the same experiences with these instructors/courses, but many of them had positive experiences as indicated by the lower means and medians. Student ratings were expected to be different since each student's experience is unique. This is why the median was calculated during the data collection process because the median essentially eliminates the outliers, both in a positive and negative sense. Examination of the median is not common in student evaluation literature.

After examining the mean question responses for each department in Table 6, Political Science represented the best scores of any one department with means of 1.80 for student learning, 1.64 instructional quality and 1.69 for instructor comparison. Philosophy, on the other hand, represented the worst scores for any one department, with means of 2.28 for student learning, 2.05 for instructional quality and 2.22 for instructor comparison. While there is little evidence in the literature exploring student evaluations ratings by department, Feldman (1978) showed Philosophy being among the courses to receive better marks on student evaluations, whereas Political Science was presented as a department in the middle of the pack in terms of student evaluation ratings. Though Feldman's (1978) findings do not align with those presented in this study, there still are too few studies to consider either one of these departmental findings generalizable.

As noted, Philosophy represented the highest percentage of general education courses in this sample, comprising 76% of its overall number of observations. Political Science, on the other hand, was on the lower spectrum of the general education courses to

total courses ratio, with general education courses only comprising approximately 19% of the total number of courses included in this study. General education courses will be discussed in more detail in Research Question #2 and might help explain the discrepancy between these two departmental ratings.

One department at the university stated in their faculty evaluation criteria that a 2.75 or better for the question on instructional quality on the SPOT survey, represented significant instructional achievement and was one of several factors included in evaluation an instructor's teaching. These faculty evaluations, which were also alluded to in the literature (McPherson & Jewell, 2007; Young, 1993; Zhao & Gallant, 2012; Cisneros-Cohernour, 2005; Arbuckle & Williams, 2003), are used to determine things like promotion, tenure, and merit bonuses. The departments used in this study, overall, had means that fell within this 2.75 threshold. In fact, there were only 136 courses that had means above 2.75 on any one question, for all 1,812 courses across 11 departments. This could mean that most the courses in this study were indeed, above average, or, that students generally rate courses and instructors better on student evaluations than the student's lived experiences. This is not something addressed in the literature, however.

The median values presented in Table 7 complimented the mean values in Table 6 as there is no median for any course above 2.00. In fact, it really puts into perspective how many courses and departments scored exceptionally well given the number of median values of 1.00, especially for question 20. Six of the 11 median values for the instructional quality question were 1.00, the best possible score. While these values are plausible, they are certainly extraordinary, and, like the argument presented in Table 6,

calls into question the validity of these evaluations as being accurate measures for course and instructor effectiveness.

In order to effectively respond to the first research question, it was important to determine the variance of the means and medians across departments. Table 8 showed the one-way ANOVA for the three SPOT questions based on department and the findings indicate that the means of these three questions are not the same across the eleven studied departments at the $p \leq .05$ level. Table 9 presents the consistent findings in examining the variance associated with the median for the three SPOT questions across the 11 departments using the non-parametric Kruskal-Wallis test. Since these tables revealed that the means and medians for the three SPOT questions were not the same across the 11 departments, the second research question was also structured by department.

The expectation, before exploring the first research question, was that student evaluation ratings would be significantly different across departments. While there were certainly differences in the ratings collected, it was not as pronounced as anticipated. In part, and as already mentioned, this could be attributed to the variety of departments used for the study, all of which were in the College of Arts and Letters. Perhaps if the study included a more diverse array of departments, with courses in Mathematics, Business, and Biology, for example, there may have been greater variation in the findings. Future studies would benefit by including a more diverse group of departments to not only examine variation in student evaluations across departments, but also across schools within the same university as well.

Question 2

The second research question for this study was: To what extent can instructor and course characteristics explain variation in student evaluations? Because it was determined that the sample collected for this study was not representative of the entire university, and the means and medians for each department were significantly different, the second research question, like the first, organized the data by department.

Faculty characteristics. The first part of addressing this question was to determine what faculty characteristics, if any, impacted student evaluations. While certain faculty characteristics were inaccessible, such as age, race, and ethnicity, other characteristics like gender, tenure status and to a certain extent, time at the university, could be determined. Using multiple regression analysis Table 10 examined faculty characteristics regression based on the means of the three SPOT questions by department. Gender was the most prominent characteristic across the 11 departments, yielding 14 significant estimated coefficients for the means of the three SPOT questions. Interestingly, these significant coefficients were mixed with nine being positive and five being negative. A negative coefficient was indicative of a better score for female instructors on student evaluations, while a positive coefficient was associated with better scores for male instructors on student evaluations.

Art, Philosophy, and Theatre courses showed no impact of gender on student evaluations. Students in courses in Anthropology, English, History, Language, and Music gave better ratings to male instructors on student evaluations. Students in Communications, Political Science, and Sociology courses displayed a tendency to give better ratings to female instructors on student evaluations. The literature presents two

findings regarding gender and student evaluations. The first is that faculty gender is not a significant factor in student evaluations (McPherson & Jewell, 2007; Liu, 2012; Krautmann & Sander, 1999), which aligns with Art, Philosophy, and Theatre courses in this study. The second is that students generally rate female instructors worse than male instructors (Wachtel, 1998; Arbuckle & Williams, 2003) and this finding was also observed in this study for courses in Anthropology, English, History, Language, and Music. No studies reviewed students rating male instructors worse than female instructors. In this study, Communications, Political Science, and Sociology students showed this tendency.

While this is certainly an interesting new finding, it is important to note that these findings are not generalizable. Additionally, it is important to consider that there are several other factors that might impact a student's perception of faculty. Smith (2009), for example, describes a correlation between faculty race and gender and their combined impact on student evaluations. Smith (2009) also suggests there are course characteristics, such as course level, that impact how students rate a faculty member of a certain race and gender. This study, unfortunately, did not have access to faculty race, but future studies should strive to include this since there is evidence that it may impact student evaluations.

Another limitation that should be presented in this discussion of the findings is that there was no student characteristics data included in this study. In terms of faculty gender, Bachen, McLoughlin, and Garcia (1999) suggest that student gender and faculty gender are correlated on student evaluations. They found that female students rated female instructors better and male instructors worse. Their study also found that male

students showed no preference. Sprinkle (2008), on the other hand, found that female students rated female faculty better, while male students rated male faculty better. Future studies would benefit by including student data to see what correlations, if any, might surface.

Length of time at the university and tenure were the other faculty characteristics collected for this study. The instructors time at the university, which, in this study, was determined by the instructor teaching at the university for five or more years or less than five years. This characteristic appeared in Table 10 of the findings 11 times for the three SPOT questions across the 11 studied departments. Despite showing up 11 times as a significant coefficient, it was only represented in five different departments:

Anthropology, English, History, Sociology, and Theatre and Dance. Like gender, it was shown as both a positive and negative significant coefficient. History and Sociology were the only two courses to show it as a negative significant coefficient, which means, for these courses, students gave faculty members better ratings if they have been at the university for five or more years. Anthropology, English, and Theatre, on the other hand, showed that students in these courses rated faculty who have been at the university for less than five years worse than their more seasoned counterparts. The number of years a faculty member has been at the university does not appear to be a significant part of the literature, but it was included in the current study because it was one of the few faculty data points that could be collected. At the very least, it was expected to be correlated with faculty tenure, which will be discussed next.

Tenure status was the final faculty characteristic collected for this study. A faculty member was considered tenured if they were, of course, tenured, or tenure-track,

as identified by their title at the time the data were collected. All other faculty were considered not-tenured or tenure track. Tenure only appeared as a significant coefficient five times for the three SPOT questions across the 11 studied departments and again, like the other faculty characteristics, it was both positive and negative. English, History, and Theatre were the only departments that showed tenure as a significant coefficient. Interestingly, all of these departments were also included for the five years at the university category. History was the only department that showed tenure as a positive significant coefficient, meaning that students in History classes for this study scored non-tenured faculty better than tenured faculty. What makes this finding even more interesting is that History was also a department where students rated instructors who had been teaching five or more years at the university, better. So, History students rated instructors who did not have tenure, but had been at the university for five or more years better than other instructors. For instructors, this is a concerning finding. From a policy perspective, it shows that non-tenured faculty score higher and continue to do so for as long as tenured faculty. It also shows that tenure is not a necessary quality for a successful instructor, which potentially could encourage administrators to offer fewer tenure positions. While the potential impact presented here might be exaggerated, this finding is nevertheless unique in that it was not seen in any other department.

English, as it turns out, had the opposite results as History. Student evaluations revealed that instructors who had been at the university for less than five years and were tenured or tenure-track, received better ratings. While still a unique finding to this study, this made more sense as a correlation, since, for tenure-track faculty, it is of the utmost importance for them to perform well in the three areas of consideration for tenure:

teaching, service, and scholarship. The literature talks at length about student evaluations being an important part of the teaching component used for recommending faculty tenure and other promotion opportunities and salary increases (McPherson & Jewell, 2007; Young, 1993; Zhao & Gallant, 2012; Cisneros-Cohernour, 2005; Arbuckle & Williams, 2003).

It is important to note that the tenure status and five-year faculty characteristics were only seen in the same department, History and English, for the same question, once. Question 19, which asked how much learning the student learned during the course, was the only question to have both characteristics present. The two other questions had either no or only one significant results for tenure and five years at the university.

These two characteristics, tenure and five or more years at the university, were expected to be correlated, since many tenure and tenure/track faculty need to be at the university for at least five years to even receive tenure. Sure enough, these were the only two faculty characteristics that yielded a positive correlation, .38, as noted in Table 12. As discussed, a correlation over .3 and less than .5 is considered a medium correlation, and, for this study, the fact that it was a medium correlation and not a large correlation between these two characteristics may be attributed to combining tenure *and* tenure track faculty. Had tenure-track faculty been combined with adjunct faculty instead, the correlation between the tenure and five year variables is expected to have been higher.

The other correlation worth mentioning from Table 12, albeit a small one, was the negative correlation between gender and tenure, -.21. This correlation suggests that female faculty are less likely to have tenure or tenure-track positions. One possible explanation for this correlation is the total number of female faculty and non-tenured or

tenure-track faculty in English, the department with the most courses included in this study. The majority of the English courses for this study were general education courses that were taught by graduate students or adjunct faculty. Nearly 60% of the English faculty included in this study were female and approximately 80% of the English faculty in this study were non-tenured/tenure-track. Since almost 30% of the total number of courses included in this study were English, and given the gender and tenure percentages above, this, in part, can help explain the negative correlation between gender and tenure. Still, there could be alternative explanations as to why this correlation, albeit small, was significant in this study. While the scope of this study does not focus on gender bias associated with faculty tenure, scholars like Winkler (2000), for example, point out that female faculty are less likely to gain tenure.

Table 11 shows the faculty characteristics regression based on the medians of the three SPOT questions by department. These findings, for the most part, were similar to those found for the means in Table 10, with some exceptions. There were no cases where the median determined a significant coefficient to be positive when the mean had it as negative and vice versa. There were only differences in the values of significant coefficients and which characteristics showed up as significant for each question. In terms of gender, Anthropology and Sociology did not have significant coefficients for the question on instructional quality, unlike in Table 10. History and Language also showed gender as a significant coefficient for instructional quality in Table 11, though Language did not for instructor comparison, like it did for means, Table 10. For tenure, English had a significant coefficient for instructional quality on Table 11, where it did not for Table 10. Conversely, History did not have a significant coefficient for tenure for instructional

quality in Table 11, like it did in Table 10. Sociology had the only difference for the five years variable by not having a significant coefficient for instructional quality in Table 11. While these differences are minor, future studies should also consider including median in their studies to compare the difference between the mean and median on student evaluations, which is something that is not addressed in the literature thus far.

While these findings may suggest a preference, future studies might benefit by including a qualitative component to help confirm the findings. In addition, as mentioned, including student data would also be advantageous to see what impact, if any, student characteristics, when combined with faculty characteristics, have on student evaluations. It could be, for example, that faculty characteristics were only significant because of student biases and class composition, which the literature describes at length (Cisneros-Cohernour, 2005; Sprinkle, 2008; Arbuckle & Williams, 2003; Platt, 1993; Galbraith, Merrill, & Kline, 2012).

Course characteristics. In order to find out what characteristics impacted student evaluations, this study ran sixty-six different multiple regressions (11 departments x 6 dependent variables). The dependent variables, in this case, were the means and medians for the three SPOT questions. Tables 13, 14, and 15 represented the linear regressions for the means of the three SPOT questions.

The course characteristic that appeared most frequently in Tables 13-15 was general education with 12 significant estimated coefficients. It appeared five times as a negative coefficient, but in only two different departments, History and Sociology. This means that student evaluations showed a preference for general education courses in History and Sociology. As a positive coefficient, General Education appeared seven

times across Anthropology, Communications, Language, and Music. For these departments, student evaluations showed that general education courses impacted ratings for the worse.

One interesting finding is that general education did not show a significant estimated coefficient for English, which had the greatest number of general education courses. This suggests that student evaluations for English courses are not impacted by a course being general education and are instead impacted by faculty characteristics and other course characteristics, day of the week and time of day. Evening English courses and English courses held on Tuesday, Wednesday, or Thursday, tended to impact student evaluations for the better while early morning courses impacted some English student evaluations for the worse.

Overall, for Tables 13-15, the following course characteristics showed up as a negative significant estimated coefficient: Fee, Division, Monday, Tuesday, Wednesday, Thursday, Friday, Afternoon, Evening, Discussion, Lab, Fall, Spring, and General Education. Course characteristics that had positive significant estimated coefficients for Tables 13-15 were: Fee, Division, Monday, Tuesday, Early AM, Evening, Fall, Spring, and General Education. These course characteristics show that there were a lot more explanations for a better rating on student evaluations than a worse rating. Across Tables 13-15, Wednesday, Thursday, Friday, Afternoon, Discussion, and Lab were the only course characteristics that appeared solely as a negative significant estimated coefficient. However, there was only one course characteristic to appear solely as a positive significant estimated coefficient, Early AM.

While some scholars conclude that time of day does not have a significant impact on student evaluations (Feldman, 1978; Wachtel, 1998; Patrick, 2011), others argue that certain times do. Koushki and Kuhn (1982), for example, found that instructors that taught classes in the early morning received lower ratings on student evaluations than instructors that taught at other times of the day. The same can be said about this study, with early morning courses showing a positive significant estimated coefficient. However, this course characteristic only showed up once across all three questions and departments, English and student learning. Does this finding, then, suggest that students who take early morning English classes do not feel that they learn as much as their peers who take the same course later in the day? Future studies would benefit by exploring this question by examining completed student evaluations and the actual grade students who completed the evaluations received in the course. It would also be beneficial to compare the student evaluation results on learning to the anticipated grade a student expected to receive in the course to see if there are any correlations. There are several studies that discuss the impact of anticipated grade on student evaluation that suggest, generally, the better the anticipated grade, the better the rating on the student evaluation (McPherson & Jewell, 2007; Feldman, 1976; Caruth & Caruth, 2013). However, these studies focus on the overall evaluation rating, not just the question on student learning, which would be something that warrants further investigation.

It is interesting to see the overlap that exists between course characteristics, with many course characteristics appearing as both positive and negative significant estimated coefficients in Tables 13-15. These course characteristics include: fee, division, Monday, Tuesday, evening, Fall, Spring, and general education. This shows, in part, that course

characteristics are not an absolute indicator of a rating on a student evaluation. There are many variables that factor into a student evaluation that may or may not be measured. Still, knowing how course characteristics can impact student evaluations will be helpful for university administrators and even faculty teaching the course.

For the most part, the significant estimated coefficient for Tables 13-15, which addressed the mean, were similar to those in Tables 16-18, which addressed the median. The biggest difference between the mean and median in examining the significance of course characteristics for the SPOT questions was the differing values. In every instance where a course characteristic and department showed a significant estimated coefficient for both the mean and median on a question, either positive or negative, the median always had the higher value if it was positive and lower value if it was negative. This finding is logical, since the median effectively eliminates outliers in student evaluations that would otherwise be present with the mean and provides a better idea of the overall rating for a course.

The median's higher positive and lower negative significant estimated coefficient, compared to its mean counterpart, did not necessarily indicate the significant estimated coefficient had a lower (more significant) p value. Most of the time, the significant estimated coefficients that showed up for both the mean and median shared the same p value, but those that differed, the mean often had the lower p value of the two. This illustrates that there is much more that accounts for significance than just a larger value.

In looking at the Tables 13-18, two things become apparent. The first is that the median had slightly different findings than the mean. As noted, future studies would

benefit by including the median when collecting student evaluation data to see if the median has significantly different findings than the mean.

The second take away is that some students do not answer every question the same. While there are some questions that have the same significant estimated coefficients, they are not all the same, nor the same values. This shows that, while learning, quality of instruction, and the instructor have a similar impact on a student's overall experience in a course, and subsequently, their evaluations, they do not impact them equally. The differing coefficients indicate that students generally rate them similarly, but not always. This illustrates the need to continue to examine multiple questions on student evaluations, not just one, which is something that scholars have already been engaging in.

Course and faculty characteristics. Up until this point, course and faculty characteristics have been examined separately in relation to their impact on the three SPOT questions in each department. Tables 19-29, however, show a regression for all course and faculty characteristics stepwise by department. Each table represents the same regression run for a different department. Also, unlike previous tables, both the mean and medians for each question are included in every table, side by side.

There were 27 different course and faculty characteristics used for Tables 19-29. One-third of these characteristics did not yield any significant coefficients in Tables 19-29. These characteristics were: instructional method, Wednesday, weekend, afternoon, lecture, discussion, lab, LLB (Bachelor of Law course), and PRF (Performance). Five of the nine insignificant characteristics were meet type, two were days in the week, and one

was time of day. One of the more interesting findings, or, rather, lack thereof, is the absence of instructional method: in-person versus online.

The literature talks at length about the way students complete evaluations, online or on-site via paper and pencil, and the pros and cons of each methodology (Hoffman, 2003; Spooren, Brockx, & Mortelmans, 2013; Stowell, Addison, & Smith, 2012; Venette, Sellnow, & McIntyre, 2010). However, there are fewer discussions about the impact of instructional method on student evaluations. Carle (2009) examined the impact of an instructor's race on student evaluations in both online and on-site classes. Carle found that an instructor's race helped to explain variance in on-site classes, but not online classes. This relationship between instructional method and a faculty member's race is just one of the many correlations that may exist between instructional method and other course and faculty characteristics.

Similar to what has already been discussed, there is not a single characteristic that explains variance in student evaluations. As much as one might like to generalize about instructional method being a major factor on student evaluations, the fact of the matter is that there is more depth to this than one might think. There are numerous explanations for why a student might prefer one instructional method over another. Scheduling, for example, is just one of the many possibilities. Perhaps a student prefers to work on their own schedule and an online course accommodates that. Or, another student might need the structure of being on campus for a course to maximize learning. In the current study, instructional method failed to explain any of the variance in student evaluation ratings.

One explanation might be the underrepresentation of online courses in the sample. There were only 246 online courses out of the 1,812 courses included in this study.

Furthermore, 233 of those online courses were in the Language department, leaving the sample with a very uneven distribution overall and within the department itself. While the findings do not show that instructional method explains variance in student evaluations, instructional method may be correlated to other course or faculty characteristics that do help explain variance in student evaluations. With this in mind, a test was run to determine if any correlations existed between instructor method and the course and faculty characteristics that showed significant estimated coefficients for Table 22 (Language). Despite not appearing as a significant estimated coefficient in Table 22, instructional method was significantly correlated to three of the seven characteristics that were listed: Friday, Fall, and SP (Special Topics). Instructional method, then, could have had an impact on these three correlated characteristics, leading to their inclusion as a significant estimated coefficient on Table 22.

Another course characteristic that did not show up in Tables 19-29 was class size. This characteristic was purposefully omitted because of the way it was input in the dataset using actual enrollment numbers. In exploring the class size variable, it became apparent that there was an uneven distribution of large and small classes. Out of the 1,812 courses used for this study, 1,610 courses had fewer than 50 students (small) and 202 courses had at least 50 students (large). A separate test was run to compare the means of the three SPOT questions in both large and small courses. Overall, smaller courses had better (lower) ratings by an average of $-.09$ on student evaluations across the means and medians for the three SPOT questions. There were only two departments where large classes had better ratings than smaller courses in the same department and

they were Sociology and Language, though it was minimal. The biggest difference between the means of small courses and large courses existed in Philosophy.

There were 74 Philosophy courses in this study. Of these 74 courses, nine were large courses and 65 were small courses. Across all small Philosophy courses, the mean of the student learning question was 2.18, the mean of the instructional quality question was 1.90, and the mean of the instructor comparison question was 2.09. Conversely, across all the large Philosophy courses, the mean of the student learning question was 3.01, the mean of the instructional quality question was 3.09, and the mean of instructor comparison question was 3.19. These are enormous discrepancies between class size in Philosophy and by far the largest of any department in this study. Even though there were only 9 large philosophy courses, it is still about 12% of the total number of Philosophy courses, which gave these subpar large class ratings enough weight to impact the department's overall rating.

In the literature, class size and the impact it has on student evaluations has been an ongoing discussion. Feldman (1978) found that class size, if it is a factor at all, has an inverse relationship to student evaluations ratings: the lower the class size the better the rating, the higher the class size the worse the rating. Others found that class size was not a significant predictor of student evaluation ratings (Liu, 2012; Jirovec, et al., 1998; Pepe & Wang, 2012). This study adds to both sides of the discussion in that, overall, the differences between the ratings of smaller classes and larger classes were minimal, but some departments, like Philosophy, displayed significant differences in ratings between small and large courses.

Instructional method and class size were two characteristics that were not in Tables 19-29, but let us now turn our attention to characteristics that were. The course and faculty characteristics that appeared most frequently in Tables 19-29 were gender, division, and general education. Of the 66 possible regressions, gender showed up in 21 of them, followed closely by division with 19 and general education with 18 appearances. It's no surprise that these course and faculty characteristics were the most prevalent in Tables 19-29, given their presence in previous tables that examined course characteristics and faculty characteristics, individually.

In an effort to consolidate Tables 19-29 to make them easier to follow, Table 30 was created, which shows the course and faculty characteristics that had at least one significant coefficient for each department, regardless of mean, median, or question number. In examining this Table, the same three course and faculty characteristics stand out as being the most prevalent across the 11 departments: gender, division, and general education. Division appeared in seven different departments, gender appeared in six different departments and general education appeared in five different departments. There was only one department that showed a positive significant coefficient for division, Sociology. The rest showed negative significant coefficients, indicating that, generally, upper division courses accounted for better ratings than lower division courses.

In the literature, there seemed to be a relationship between course level and student evaluation rating: the higher the course level, the better the rating (Feldman, 1978; Wachtel, 1998; Patrick, 2011). This study affirms this notion; six of the seven departments that yielded a significant estimated coefficient for division were negative, indicating upper division courses accounted for better ratings on student evaluations.

Gender appeared in six different departments, both as a positive and negative coefficient, four times and two times, respectively. These results are similar to that of Table 10, which had gender appear in seven different departments, where four departments had positive coefficients and three departments had negative coefficients. The only different department between these two tables was Sociology, which had a negative coefficient in Table 10. Though gender was already addressed in detail previously, it is important to acknowledge that it still exists as a predictor for explaining variance in student evaluations when measured alongside course characteristics.

General education was the third largest characteristic and showed up for six different departments across Tables 13-18: Anthropology, Communications, Language, Music, History, and Sociology. Table 30 had a few differences. Anthropology and Sociology did not have a significant coefficient for general education, like they did in Tables 13-18. Philosophy and English, on the other hand, had significant coefficients in Table 30, but did not in Tables 13-18. Communications, History, and Music were the only departments to have general education as a significant coefficient in Tables 13-18 and Table 30.

The biggest standout in general education is its presence as a significant coefficient for English in Table 30. As noted, English had the most general education courses of any department used in this study, but for Tables 13-18, failed to show it as being a significant coefficient. English, however, showed general education as a positive significant coefficient in Table 30. Communications, Music, and Philosophy also showed general education as a positive coefficient and History was the only negative significant coefficient for general education in Table 30.

While there are some scholars examining the impact of required versus elective courses on student evaluations (Pepe and Wang, 2012; Feldman, 1978; Cisneros-Cohernour, 2005), there is still much more research to be done on this topic. This study supports the existing literature by confirming that general education courses at this university, for the most part, when significant, accounted for less favorable student evaluation ratings. However, as there are with most course and faculty characteristics examined in this study, there are exceptions. In the case of general education in Table 30, History showed these courses to impact student evaluation ratings for the better, which is something that was previously absent in the literature.

Class time is something that hasn't been discussed in too much detail in this study and warrants further investigation given its prevalence in the literature (Feldman, 1978; Wachtel, 1998; Patrick, 2011; Koushki & Kuhn, 1982). Most scholars found that the time a course takes place does not explain variance in student evaluation ratings (Feldman, 1978; Wachtel, 1998; Patrick, 2011). Others found that certain times impact student evaluation ratings, such as early morning classes, which typically receiving lower ratings (Koushki & Kuhn, 1982). English, Language, and Music in this study showed similar findings, with early morning courses in these departments impacting student evaluations for the worse.

Late morning courses also displayed significant coefficients in Table 30, both positively and negatively, in Philosophy and Music, respectively. Evening courses were the only class time to always account for better student evaluation ratings. Still, there were only two departments that showed evening as a negative coefficient, Language and

Political Science. Afternoon doesn't appear to explain variation in any of the 11 departments used for this study.

Like other course and faculty characteristics, class time does not show up in every department as a factor in explaining variance in student evaluation ratings. This does not mean, though, that class time, or the other course and faculty characteristics, are irrelevant for departments where they fail to show a significant coefficient. On the contrary, it is perhaps only that other characteristics are more significant and in effect, making something others less significant, as a result. It is for this reason that all of the course and faculty characteristics used in this study, should continue to be studied, for it is not known how they might impact student evaluations at other universities or in other departments.

There are no two departments that shared the exact same course and faculty characteristics as significant coefficients as shown in Table 30. This suggests that each department is unique, and the course and faculty characteristics findings for each department are not generalizable; what is true for one department is not necessarily true for another. Therefore, course and faculty characteristics and the impact they have on student evaluations should continue to be studied in the future.

Some departments had more significant coefficients than others. Music, for example, had the most significant characteristics, eight. These eight characteristics were evenly split as positive and negative significant coefficients in Table 30. The department with the fewest different course and faculty characteristics for significant coefficients was Communications, where only general education and Gender were significant coefficients. This department was also evenly split with positive and negative significant coefficients.

There was only one department that failed to show both a positive and negative coefficient for course and faculty characteristics, Political Science.

This department had six different course and faculty characteristics that all yielded negative significant coefficients in Table 30. Of any of the departments, this one makes the most sense since it was Political Science that had the best overall ratings on Questions 19, 20, and 21. Still, it is unusual that there was not one course or faculty characteristic that had a positive significant coefficient, since there was variation in the student evaluation ratings; not every Political Science course got the same rating. There were some courses that had worse scores than others, but apparently, a poorer rating cannot be explained by any of the course and faculty characteristics used for this study. Perhaps there were other course and faculty characteristics not collected in this study that accounted for worse student evaluation ratings in Political Science. Future studies should attempt to collect additional course and faculty characteristics to see if they too help explain variation in student evaluations.

In every department except Theatre and Dance, as noted in Table 30, there was at least one course characteristic and one faculty characteristic for each department that had a significant coefficient. Theatre and Dance only had significant coefficients for course characteristics, two of which were negative (Fall and Studio) and one being positive (Spring). Does this suggest that faculty characteristics had no impact on student evaluations? No. Referring back to Tables 10 and 11 that focused solely with faculty characteristics, tenure and five years at the university both had significant coefficients for Theatre and Dance. Still, since these same characteristics were not present in the

combined course and faculty characteristics Table (30), it suggests that the impact was less significant than the course characteristics that produced significant coefficients.

While some course and faculty characteristics yielded both positive and negative significant estimated coefficients depending on the department, these characteristics never had both positive and negative significant estimated coefficients for one department. This, again, supports the notion that the three questions used for this study are correlated and student responses don't vary a large amount from question to question. For the most part, up to this point course and faculty characteristics have been grouped together in discussing the three SPOT questions. While these three questions are broad and somewhat related, they are nevertheless unique questions. Since they are unique, they do not necessarily all have the same responses from an individual student. The following section will discuss several individual course and faculty characteristics and the extent to which they yielded (or did not yield) a significant coefficient for a particular SPOT question in a department.

The relationship of course and faculty characteristics with individual questions and departments. In taking a closer look at Tables 19-29, specifically, what course and faculty characteristics were significant for one question, but perhaps not others, there were several findings worth discussing. The first was with the question on student learning. In three of the four departments where gender was significant in terms of student learning gender had a positive significant coefficient, indicating that students felt that a male instructor was more effective with regard to student learning. However, for the student learning question, gender yielded fewer significant coefficients than the other two questions, instructional quality and instructor comparison. Interestingly, instructor

comparison had the most significant coefficients across the 11 departments for gender. Four of the six departments that yielded significant coefficients for gender on the instructor comparison question were positive. This is an interesting finding because the question asks students to compare this instructor to others they have had. While male instructors typically scored better here, we do not know the gender of the other instructors the student had previously and how they were rated.

Division was one of the few characteristics to stand out when looking at individual questions. As it turns out, division had twice as many significant coefficients for the instructor comparison question than the other two questions. The instructor comparison question for division yielded ten significant coefficients compared to five for instructor quality and five for student learning. Out of the ten significant coefficients for division on the instructor comparison question, only one of them was positive. This is telling of the relationship between course level and how students rate instructors. Instructors who teach upper division courses have a major advantage of being rated better than instructors who are teaching lower division courses. Again, what we do not know is what courses or instructors these students previously had that may help explain this variation in student evaluation ratings.

Overall, these were how the significant coefficients broke down by the mean and median for each question. The mean of the student learning question had 24 significant coefficients across the 11 departments while the median of the student learning question had 17 significant coefficients across the 11 departments. The mean of the instructor quality question had 22 significant coefficients and the median of the instructor quality question had 28 significant coefficients. Lastly, the mean instructor comparison question

had 27 significant coefficients across the 11 departments and median of the instructor comparison question had 28 significant coefficients. Looking at the means and medians as a whole, the three SPOT questions had an equal number of significant coefficients, 73 for the mean and 73 for the median.

The expectation was that there would be significant differences between departments, questions, and characteristics. However, the biggest differences were between departments; not all departments had the same course and faculty characteristics as significant coefficients. Still, future studies should consider the impact of course and faculty characteristics on individual questions to see what nuances may exist.

Discussion Summary

This study supported the literature in several ways. It confirmed certain findings that already existed in the literature such as: upper division courses get better ratings than lower division, early morning courses get worse ratings, general education courses get worse ratings, and female instructors get worse ratings than male instructors. Of course, these confirmed findings are general and were not prevalent in every department, they were only found in at least one department.

This study also added to the literature in several ways. It found for some departments that: gender can be insignificant, students can rate male instructors worse than female instructors, an instructors time at the university and tenure can both make ratings better and worse on student evaluations, general education courses can get better ratings, and evening courses get better ratings. Again, these contributions only had to be true for one department; this does not represent something that was found in every department and it is not generalizable.

Many of these additions can be attributed to having a diverse sample of departments. Few studies in the literature looked at more than one department, so it was interesting to see the diversity of the findings that were presented, as a result.

While there were many things this study did find, there were many things that it also did not find. Something that was expected to be significant was the difference in student evaluation ratings between online and onsite courses across departments. However, the data collected did not yield enough diversity of course type across departments to make it significant. This study also did not find any single characteristic that always predicts student evaluations to be rated a certain way. The vast majority of the course and instructor characteristics in this study were two-dimensional: they impacted student evaluations in a positive or a negative way, depending on the department.

Overall, the courses used for this study were rated above average by students, as indicated by their means and medians for the three SPOT questions. The median had more pronounced significant coefficients than the mean, indicating that the median explained more variation in student evaluations than the mean. Though the differences between the means and medians of different departments was minimal, it was nevertheless, apparent. Future studies might find that including departments from different colleges at a university might show a greater variance in student evaluation ratings than was evident in this study.

Implications for Policy

From a higher education policy perspective, the findings presented in this study might have some implications. One of the bigger concerns is with faculty tenure. While

tenure itself was not overly significant variable in the study, tenure-track faculty have a lot at stake with student evaluations given their weight in decisions on tenure (Xu, 2012; Shirbagi, 2011; Spooren, Brockx, & Mortelmans, 2013; Young, 1993; Cisneros-Cohernour, 2005). For example, giving a tenure-track faculty member only early morning, lower division, general education courses may be setting them up to underperform on student evaluations compared to their colleagues who teach evening, upper division, elective courses. This study shows that student and faculty characteristics do help to explain variation in student evaluation ratings and, as such, student evaluations should be examined closely both within and across departments at the university. Administrators should consider conducting a similar study at their university to determine what, if any, student biases exist at their university.

While many of the course characteristics can be addressed by evenly distributing courses amongst faculty (equal number of morning and evening classes, or, equal number of required and elective courses, for example), there are other characteristics that cannot be addressed, such as faculty gender. Both the literature and this study show that faculty gender can help to explain variance in student evaluations. While absent in this study, the literature suggests that the same might also be said about a faculty member's race and ethnicity. These faculty characteristics should be explored within the administrator's own university to see if this is evident and, if necessary, weight evaluations accordingly.

Another consideration to administrators should be the distribution of money to their departments or schools from the provost. What criteria does the provost use for distribution? If even a portion of the criteria involves student evaluations in some capacity, then it could be advantageous to see where the administrator's school or

department ranks across the entire university. If more money is allocated to departments that receive better ratings, then it is up to the administrator to make sense of the ratings. Not all departments are rated equally. If these ratings are a result of student bias, for example, then administrators should petition to have their department or school's evaluations weighted, accordingly.

While this study is by no means a generalizable guide for all universities, it hopefully makes administrators more aware of the potential things to look out for and how course and faculty characteristics may be impacting student evaluations at their universities.

Limitations

There were a lot of things that this study had in its favor including, but not limited to: a large, public dataset that spanned multiple years, inclusion of the median in addition to the mean, 11 different departments, both course and faculty characteristics, and focusing on one school within one university. These benefits, however, were not without significant limitations.

While it could be viewed as a benefit to have studied one school at one university, it was also a limitation. The findings are not generalizable. In addition, while there were 11 different departments, they were all within the same school. Future studies would benefit by exploring multiple schools and multiple universities to have a more diverse sample in hopes of having findings that are more generalizable.

This study also only examined undergraduate courses. While this is not uncommon from the literature, future studies would benefit by examining both graduate and undergraduate courses to see if there are any differences in student evaluation ratings.

This study was also limited in terms of accessibility. Since this study used a public dataset, there were certain course and faculty characteristics, such as faculty member's race and ethnicity, which were not accessible. In addition, it was not possible to collect any student data, which the literature argues, can be an important component in explaining variance in student evaluations. Things like gender, race and ethnicity, major, and anticipated grade were not able to be collected. Future studies are encouraged to gather student data to see how they impact student evaluations and how they are correlated to course and faculty characteristics.

This study could have also benefitted from the use of qualitative methods. Interviews with students, specifically, could have given them the opportunity to explain their student evaluation ratings and allowed the researcher to compare these comments to the actual ratings. Future studies could benefit by including a mixed methodological approach.

The distribution of courses was also a limitation in that the courses were not evenly distributed across a number of variables. As noted previously, English and Language alone comprised more than 50% of this study. A more even distribution of courses for each department would be important for future studies, though not at the cost of eliminating other selection criteria, such as the two-thirds threshold or minimum number of responses.

Implications for Future Research

While there has already been some mention of the implications for future research, this section will discuss additional considerations for future research and elaborate on a few already mentioned. The purpose of this study was to determine what

course and faculty characteristics, if any, helped to explain variation in student evaluations. While this study accomplished this to an extent, future studies would benefit by diversifying and balancing their sample, in addition to collecting student data as well.

The literature would benefit from a study that is not limited in terms of accessibility, so every course, faculty, and student characteristic the researcher wants, can be included. An additional course characteristic that might be worth including that is seemingly absent in the literature is the number of students that dropped a particular course – which may have an impact on student evaluation ratings. Additional faculty characteristics that should be used in future studies include, but are not limited to: race and ethnicity, sexual orientation, and age. Much of the literature is devoted to the impact of student characteristics on student evaluations, since some student characteristics have been used to help explain variation in student evaluations in previous studies. The absence of student characteristics in this study was unfortunate, since these three characteristics can be significantly correlated.

The literature would also benefit greatly if future studies included several schools, within several universities. The more universities, schools, and courses that can be included in a study on student evaluations, the better, and, the more generalizable the findings will become.

One of the challenges researchers may encounter when examining student evaluations at multiple universities is the difficulty in finding a standardized evaluation. It is often hard enough to find a standard evaluation that is used across one university, let alone multiple universities. Still, most student evaluations have similar questions and

scoring and even if the researcher studies universities that use different evaluations, there are still ways to statistically compare the data.

In addition, future studies should consider the impact of course and faculty characteristics on individual questions. While grouping questions for analysis can be important, examining individual questions, and their significant course and faculty characteristics, will provide deeper findings.

There is still much to be learned from student evaluations and the impact course, faculty, and student characteristics have on them. While we know that there is not any one characteristic that impacts ratings the same way, every time, for every department, we also assume there are some correlations that exist that have yet to be discovered. The more studies there are on student evaluations, the closer we will be to discovering their nuances, which will give us a better idea of how to maximize their utility.

REFERENCES

- Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, 13(2), 153-166.
- Allan, J., Clarke, K., & Jopling, M. (2009). Effective teaching in higher education: Perceptions of first year undergraduate students. *International Journal of Teaching and Learning in Higher Education*, 21(3), 362-372.
- Amin, M. E. (1994). Gender as a discriminating factor in the evaluation of teaching. *Assessment and Evaluation in Higher Education*, 19, 134-143.
- Arbuckle, J. & Williams, B. D. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles*, 49, 507-516.
- Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education*, 48(3), 193-210.
- Basow, S. A., Codos, S., & Martin, J. L. (2013). The effects of professors' race and gender on student evaluations and performance. *College Student Journal*, 47(2), 352-363.
- Brockx, B., Spooren, P., & Mortelmans, D. (2011). Taking the grading leniency story to the edge: The influence of student, teacher, and course characteristics on student evaluations of teaching in higher education. *Educational Assessment, Evaluation and Accountability*, 23(4), 289-306.
- Campbell, H. E., Gerdes, K., & Steiner, S. (2005). What's looks got to do with It? Instructor appearance and student evaluations of teaching. *Journal of Policy Analysis and Management*, 24(3), 611-620.
- Carle, A. C. (2009). Evaluating college students' evaluations of a professor's teaching effectiveness across time and instruction mode (online vs. face-to-face) using a multilevel growth modeling approach. *Computers & Education*, 53(2), 429-435.
- Caruth, D. L., & Caruth, G. D. (2013). Grade inflation: An issue for higher education? *Turkish Online Journal of Distance Education*, 14(1), 102-110.
- Cashin, W. E. (1995). Student ratings of teaching: The research revisited. IDEA Paper No. 32. Retrieved from <http://0-search.ebscohost.com.sally.sandiego.edu/login.aspx?direct=true&db=eric&AN=ED402338&site=ehost-live>

- Centra, J. A. (1993). Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness. *The Jossey-Bass Higher and Adult Education Series*. Retrieved from <http://0search.ebscohost.com.sally.sandiego.edu/login.aspx?direct=true&db=eric&AN=ED363233&site=ehost-live>
- Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, 44(5), 495–518.
- Cisneros-Cohernour, E. J. (2005). Validity and evaluations of teaching in higher education institutions under positivistic paradigm. *Quality of Higher Education*, 2, 84–95.
- Erdle, S., Murray, H. G., & Rushton, J. P. (1985). Personality, classroom behavior, and student ratings of college teaching effectiveness: A path analysis. *Journal of Educational Psychology*, 77(4), 394–407.
- Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education*. Retrieved from <http://0search.ebscohost.com.sally.sandiego.edu/login.aspx?direct=true&db=eric&AN=EJ132876&site=ehost-live>
- Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education*, 9, 199–242.
- Feldman, K. A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education*, 139.
- Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education*, 53(3), 353–374.
- Harris, T., & Twinaime, L. (2012). Student evaluations of teaching: The students' perspective. *Studies in Learning, Evaluation, Innovation & Development*, 9(1), 145–156.
- Hoffman, K. M. (2003). Online course evaluation and reporting in higher education. *New Directions for Teaching & Learning*, 2003(96), 25–29.
- Jirovec, R. L., Ramanathan, C. S., & Alvarez, A. R. (1998). Course evaluations: What are social work students telling us about teaching effectiveness? *Journal of Social Work Education*, 34(2), 229–36.

- Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality*, 64(2), 311-337.
- Krautmann, A. C., & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review*, 18(1), 59-63.
- Kyei-Blankson, L., & Nur-Awaleh, M. (2010). An examination of faculty effectiveness in technology integration in teaching from students' perspectives. *International Journal of Learning*, 17(6), 451-460.
- Liu, O. (2012). Student evaluation of instruction: In the new paradigm of distance education. *Research in Higher Education*, 53(4), 471-486.
- Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, 52, 77-95.
- Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76(5), 707-754.
- McPherson, M. A., & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly (Wiley-Blackwell)*, 88(3), 868-881.
- Murray, H. G. (2005). Student evaluation of teaching: Has it made a difference? Paper presented at the annual meeting of the *Society for Teaching and Learning in Higher Education*, Charlottetown, Prince Edward Island.
- Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology*, 82(2), 250-261.
- Norvilitis, J. M., & Zhang, J. (2009). The effect of perceived class mean on the evaluation of instruction. *Educational Assessment, Evaluation and Accountability*, 21(4), 299-311.
- Patrick, C. L. (2011). Student evaluations of teaching: Effects of the big five personality traits, grades and the validity hypothesis. *Assessment*, 36(2), 239-249.
- Pepe, J. W., & Wang, M. C. (2012). What instructor qualities do students reward? *College Student Journal*, 46(3), 603-614.
- Platt, M. (1993). What student evaluations teach. *Perspectives on Political Science*, 22(1), 29-40.

- Shirbagi, N. (2011). Iranian university teachers' and students' views on effectiveness of students' evaluation of teaching. *Quality of Higher Education*, 8, 118–131.
- Smith, B. P. (2009). Student ratings of teaching effectiveness for faculty groups based on race and gender. *Education*, 129(4), 615–624.
- Smith, G., & Anderson, K. J. (2005). Students' ratings of professors: The teaching style contingency for Latino/a professors. *Journal of Latinos & Education*, 4(2), 115–136.
- Sonntag, M. E., Bassett, J. F., & Snyder, T. (2009). An empirical test of the validity of student evaluations of teaching made on ratemyprofessors.com. *Assessment & Evaluation in Higher Education*, 34(5), 499–504.
- Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598–642.
- Sprinkle, J. E. (2008). Student perceptions of effectiveness: An examination of the influence of student biases. *College Student Journal*, 42(2), 276–293.
- Stowell, J. R., Addison, W. E., & Smith, J. L. (2012). Comparison of online and classroom-based student evaluations of instruction. *Assessment & Evaluation in Higher Education*, 37(4), 465–473.
- Thyer, B. A., Myers, L. L., & Nugent, W. R. (2011). Do regular social work faculty earn better student course evaluations than do adjunct faculty or doctoral students? *Journal of Teaching in Social Work*, 31(4), 365–377.
- Venette, S., Sellnow, D., & McIntyre, K. (2010). Charting new territory: Assessing the online frontier of student ratings of instruction. *Assessment & Evaluation in Higher Education*, 35(1), 97–111.
- Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, 23(2), 191.
- Xu, Y. (2012). Developing a comprehensive teaching evaluation system for foundation courses with enhanced validity and reliability. *Educational Technology Research & Development*, 60(5), 821–837.
- Young, R. D. (1993). Student evaluation of faculty: A faculty perspective. *Perspectives on Political Science*, 22(1), 12–16.
- Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education*, 37(2), 227–235.

APPENDIX:

SAMPLE SPOT EVALUATION

Note: the instructor's name, course, and campus were omitted from this sample. In addition, you will notice the numbers skip from 15 to 19 on the evaluation. This is how it actually appears in the online database. Questions 16-18 were expected to be open-ended questions.

Enrolled:	27
Responses:	23
Percent:	85.2

	Completely Agree (1)	Somewhat Agree(2)	Somewhat Disagree(3)	Completely Disagree(4)	No Response	Mean
1. Clearly stated the objectives of the course.	73.9	21.7	4.3	0.0	0.0	1.30
2. Covered what was stated in the course objectives.	65.2	21.7	8.7	4.3	0.0	1.52
3. Was organized and prepared for class	69.6	17.4	8.7	0.0	4.3	1.36
4. Communicated ideas effectively	78.3	8.7	13.0	0.0	0.0	1.35
5. Told students how they would be evaluated	69.6	17.4	13.0	0.0	0.0	1.43
6. Gave assignments that assisted in learning the material	73.9	17.4	4.3	4.3	0.0	1.39
7. Gave useful feedback on coursework	47.8	34.8	17.4	0.0	0.0	1.70
8. Made the subject interesting	69.6	17.4	13.0	0.0	0.0	1.43
9. Encouraged students to think critically	56.5	21.7	21.7	0.0	0.0	1.65
10. Was willing to listen to students' questions and opinions	73.9	17.4	8.7	0.0	0.0	1.35
11. Was available during office hours or appointment times	82.6	8.7	8.7	0.0	0.0	1.26
12. Showed respect for students	91.3	4.3	4.3	0.0	0.0	1.13
13. Was concerned with whether students learned	73.9	17.4	8.7	0.0	0.0	1.35
14. Was interested in teaching	82.6	13.0	4.3	0.0	0.0	1.22
15. Gave exams that reflected the material covered*	78.3	17.4	4.3	0.0	0.0	1.26

*If no exams were given in this course, leave blank.

	An exceptional amount(1)	More than usual(2)	About as much as usual(3)	Less than usual (4)	Almost nothing(5)	No response	Mean
19. How much do you think that you have learned in this course?	26.1	30.4	21.7	17.4	4.3	0.0	2.43
	Excellent(1)	Very Good (2)	Good(3)	Fair(4)	Poor(5)	No response	Mean
20. Rate the quality of instruction as it contributed to your learning in the course.	52.2	21.7	17.4	8.7	0.0	0.0	1.83
	One of the most effective (1)	More effective than others(2)	About as effective as others(3)	Less effective than others(4)	One of the least effective(5)	No response	Mean
21. What is your rating of this instructor compared to other instructors you have had?	43.5	26.1	21.7	4.3	4.3	0.0	2.00