ANALYZING AND CONTROLLING BIASES IN STUDENT RATING OF INSTRUCTION

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Yue Zhou

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Program:
Applied Statistics

November 2018

Fargo, North Dakota

North Dakota State University
Graduate School

Title

ANALYZING AND CONTROLLING BIASES IN STUDENT RATING OF INSTRUCTION

By

Yue Zhou

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Gang Shen
Chair

Dr. Megan Orr

Dr. Guodong Liu

Approved:

January 7, 2019	Dr. Rhonda Magel
Date	Department Chair

ABSTRACT

Many colleges and universities have adopted the student ratings of instruction (SROI) system as one of the measures for instructional effectiveness. This study aims to establish a predictive model and address two questions related to SROI: firstly, whether gender bias against female instructors at North Dakota State University (NDSU) exists and, secondly, how other factors related to students, instructors and courses affect the SROI. In total, 30,303 SROI from seven colleges at NDSU for the 2013-2014 academic year are studied. Our results demonstrate that there is a significant association between students' gender and instructors' gender in the rating scores. Therefore, we cannot determine how the gender of an instructor effects the course rating unless we know the composition of genders of students in that class. Predictive proportional odds models for the students' ordinal categorical ratings are established.

ACKNOWLEDGEMENTS

I would like to express my most sincere appreciation to my advisor Dr. Gang Shen for his help, advice, patience, encouragement and all the supports throughout the entire process. He is an awesome advisor that gave me insight on the scientific community especially for statistics. I would never have finished my thesis without his guidance.

I want to thank the members of my committee. Dr. Megan Orr and Dr. Guodong Liu for their help and feedback on my thesis.

A special thanks goes to Dr. Curt Doetkott for sharing the dataset with me.

I also wish to thank my friends and other graduate students: Juechen Yang, Meiyan Sun, Yun Zhou, Qiang Li, Yu Sun, Deling Zhang and Gunjan Gugale for their support.

Finally, I want to thank my parents Guolin Zhou, Yunhua Li and my sister Shuang Zhou for their unquestioning support, sacrifice and love. They are the most important people in my life.

I am truly grateful for all that you did to help me in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

A student’s academic performance mainly depends on his or her background, effort and time; however, this performance is also affected by the instructor's teaching ability. An instructor imparts professional knowledge and resolves doubts. Thus, instructors are key elements of the entire higher education system. Student ratings of instruction (SROI) is a convenient metric for faculty teaching performance assessment and has been widely adopted for years in American institutions. However, SROI’s validity, fairness, and effectiveness have long been questioned, particularly, the issues of bias against female instructors. Recently, there have been calls for institutions to stop giving an inordinate amount of weight to student evaluations when making employment decisions until the biases can be accounted for, addressed, and eliminated. Unfortunately, there’s no consensus on how best to eliminate these biases. In this work, we establish a proportional odds model for addressing and controlling the biases for SROI.

As early as the 1960s, American universities and colleges began to utilize informal SROI. Since then, this measurement has been used for academic personnel performance evaluation and curriculum quality assessment because it provides direct and quick feedback. Currently, SROI is considered one of the main methods of judgement under the promotion-and-tenure category of teaching. However, this assessment’s potential biases are often subject to debate. In addition to the widely concerned gender bias, multiple factors, such as class size, teaching environment, clarity of expression, interaction with students, and classroom activities, have also been found to impact student evaluation. The course type and the amount of knowledge perception also caused student ratings to differ significantly. Therefore, instead of using SROI as the sole, definitive, and objective measure of teaching quality, it would be wiser to explore an approach to justify the assessment results with reasonable control of bias.

In this study, a statistical model to accurately explain and predict the relationship between student satisfaction and various related factors in the curriculum was established using the SROI from 30,303 students in seven colleges collected during the academic year 2013-2014 at North Dakota State University (NDSU). At NDSU, students could evaluate instructors' performance by completing a questionnaire at the end of the semester. Questions about the instructor and course quality were provided to students for evaluation. Additional information about the student and instructor, such as gender and college, were also included on the form. All the forms were scanned and stored in the university database, and the data were protected and regulated by the Institutional Review Board (IRB). Each question had categories from 1 to 5, respectively representing "Very Poor/Strongly Disagree," "Poor/Disagree," "In Between/Neutral," "Good/Agree," and "Very Good/Strongly Agree."

Only fully completed questionnaires were included in this study. Within the dataset, an instructor may teach multiple courses, and a student may enroll in and rate multiple courses. The five categories for each question were treated as ordinal data and evaluated using proportional odds cumulative logit model. In our model, the rating of an instructor's performance by each individual student will be analyzed associated with the demographic information of instructor and student, and the information of the class. Base on this, the effects of gender and the gender interaction are studied incorporate all effects of related covariates. To better assess the impact of SROI, we established two types of models based on the proportional odds cumulative logit model method of analysis for two dependent variables ("Instructor as a teacher" and "Quality of this course") respectively. For the first response variable, one type of models uses the class information, and the other type of models uses class information and instructor's performance. For the second response variable, one type of models uses class information and student's performance, and the other type of models uses all the information that can be obtained from the SROI. Backward elimination is used for optimizing models. In addition, goodness of fit and the accuracy of models are evaluated in our study.

2. LITERATURE REVIEW

Delivering high quality teaching is one of the most pivotal missions of universities around the world. A variety of measurements have been adopted to foster the improvement of teaching quality, including department chair and colleague rating, instructor self-assessment, and student rating (Bowles, 2000).

In 1998, Huemer indicated that SROI is a reliable way to evaluate instructors' performance. He also reported that ratings by colleagues are not reliable as they have no main agreement with other observers for instructor ratings. Therefore, the evaluation of teaching by students is highly valued by academic institutions. Student ratings have been used to continuously improve the quality of teaching and learning. Most studies indicated that instructor's tenure, promotion, and salary are also potentially affected by the student rating (Punyanunt & Carter, 2017; Whitworth, Price, & Randall, 2002). In addition, student satisfaction is important for the reputation and future enrollment of higher education institutions (Long, Ibrahim, & Kowang, 2015). However, the reliability of student evaluations has become less trusted over the years because different biases have been identified to complicate their interpretation. More evidence has been discovered that student evaluations of instruction are often biased. In 2017, Hornstein claimed that SROI is mainly used for evaluating the performance of faculties but SROI involves biases which made faculty under pressures. This is common especially for tenure-track faculty as the tenure system is based on merit at most universities. Therefore, Hornstein insisted that SROI are not an adequate assessment for summative evaluation of faculty. The value of student evaluations is controversial. The research of the role of gender bias in student evaluations can be traced back to the 1980s (Basow & Silberg, 1988). Another recent study confirmed that the student evaluations are biased against women (Mitchell & Martin, 2018). Mitchell reported that in SROI, the language used to evaluate male instructors is significantly different than female instructors. Whether students rate male and female instructors differently, even when those instructors performed the same, has been explored by many researchers (Maricic, Djokovic, & Jeremic, 2016). Boring et al. found that female instructors were treated with gender biases in SROI (Boring et al., 2017). In the same year, Boring (2017) used the logit regression and fixed effects model to analyze the possible gender biases in SROI for a French University. She concluded that male students favor male instructors even nothing can prove that male instructors are better than female instructors.

MacNell (2015) performed an online experiment to explore the gender bias in SROI. In the experiment, each instructor used two different genders to teach the same online course. Students did not know the instructor's real gender. He concluded that regardless of the instructor's actual gender, the male instructors received significantly better scores than female instructors.

Moreover, Rosen (2017) indicated that the significant difference in ratings between the male and female instructor also depends on the teaching discipline. Gender bias may not exist in all disciplines because female instructors received similar scores to male instructors in some fields, such as chemistry, while they received less satisfactory scores in other disciplines, such as history. In addition, students assigned lower scores to both male and female instructors who taught science and engineering than those who taught arts and humanities.

In contrast, Bachen, Mcloughlin and Garcia (1999) analyzed the influence of gender schema on students' perceptions and ratings of male and female instructors from the psychological perspective. They found that the relationship between the student's gender and the instructor's gender was significant. Female instructors received relatively higher scores than male instructors from the female students, while no significant difference in rating was found for the male students. Meyer confirmed Bachen's results, indicating there is a significant relationship between instructors’ gender and students’ gender (Centra & Gaubatz, 2000; Meyer, Doromal, Wei, & Zhu, 2017). Whitworth (2002) studied the effect of faculty’s gender on student ratings and found that female instructors received better scores than male instructors. Likewise, Maricic (2016) indicated that it is more common for female faculty to receive higher ratings. Smith et al. (2007) used a large sample size to research the gender influence on student ratings of instructors and claimed that female instructors received better ratings from both female and male students.

Unfortunately, gender is not the only type of bias present on student evaluations. The validation of evaluation, the effect of class size, and course type, as well as other factors, are commonly studied biases.

Effectively designing proper questions in student evaluation is a controversial topic (Marsh & Bailey, 1993). A valid student evaluation form should cover key dimensions of evaluating teaching effectiveness (Dodeen, 2013). Dodeen claimed that when constructing an effective form to assess teaching quality, many characteristics of instructors’ performance and classroom environment must be considered: learning, fairness, objectivity, interaction with students, clarity, teaching methods, effective feedback, grading, and high standards. In addition, many other factors have been revealed to have an influence on the accuracy of the assessment, including class size (Bennett, 1982; Jones, 2017; Smith et al., 2007) course type (required versus elective) (Feit, 2014), class level (Feit, 2014; Whitworth et al., 2002), college (Bennett, 1982; Feit, 2014), gender (Dodeen, 2013; MacNell et al., 2015; Punyanunt & Carter, 2017; Rosen, 2017), and expected grade (Griffin, 2006). The SROI of NDSU was designed to cover the major aspects evaluated in these studies.

Similar to gender bias, the effect of class size, course type, and other factors are biases on SROI. Jones (2017) indicated that the size of a class will influence the students' perceptions. He found that the instructors received lower ratings when the class size was large. In addition, Ibrahim (2011) found that the influence of increasing class size is greater than increasing the number of factors on the generalizability coefficients. The relationship between class size and the instructor’s gender was studied by Smith (2007), who found that male instructors are more likely to teach courses with larger class sizes. Class level is another bias that exists on SROI. Whitworth et al. (2002) reported that graduate students have higher satisfaction with quality of the teaching than undergraduate students. Furthermore, the role of discipline and course type were studied by researchers. Feit (2014) claimed there is a significant effect of discipline and course type on SROI; however, class level does not play an important role. He also found that students in the STEM disciplines, such as science, engineering and mathematics, tended to have the lowest satisfaction with the quality of teaching, whereas instructors in educational disciplines received the highest scores.

3. DATA AND METHODOLOGY

3.1. Research Objective

The purpose of this study is to determine whether there is a gender bias against male or female instructors at NDSU and to establish a statistical model to accurately explain how the satisfaction of students is related to the following cofactors: the performance of the instructor, gender, class level, course type, expected grade, and college.

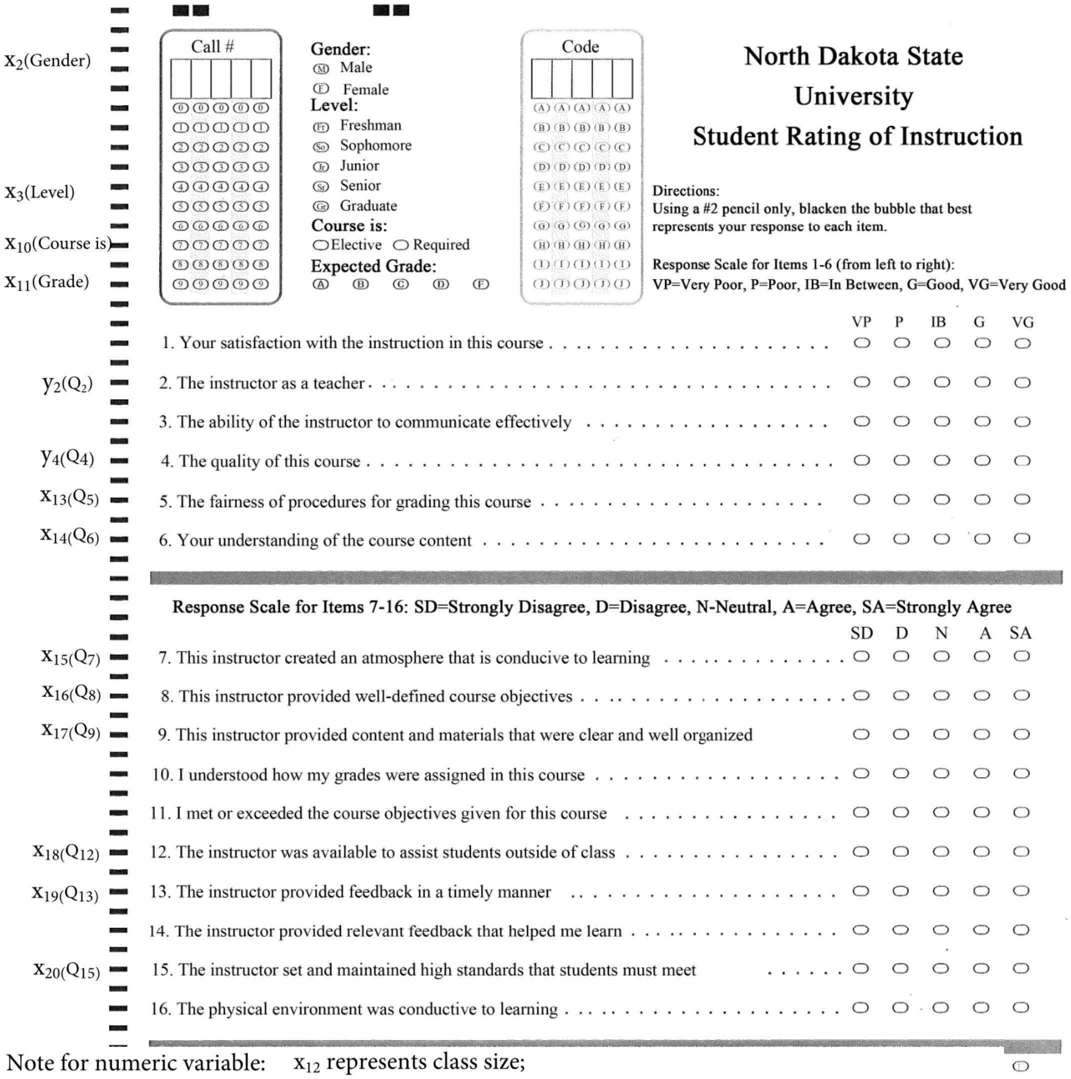
3.2. Data Collection

In total, 30,303 SROI from seven colleges during the 2013-2014 academic year at NDSU were collected. Sixteen evaluative questions about the courses and instructors were asked to students and are presented in the Figure 3.1, below.

We are interested in the variables' effects on the SROI and the course quality. Question 2 and question 4 were treated as response variables. Eight questions (questions 5 through 9 and questions 12, 13, and 15) were selected for establishing models.

The influence of seven more variables related to course, students' and instructors' information were also considered during our study. First five variables are indicator variables. The first variable was the instructor's gender. The data was coded as 1 for male and 0 for female. Overall, 18,049 instances of student feedback for male instructors and 12,254 for female instructors were collected. One instructor may have taught different courses with different class sizes or in different colleges.

The second variable was the student's gender. Again, the data was coded 0 for female and 1 for male. Instructor evaluations by 14,782 female students and 15,621 male students were collected. Class level was the third variable: 29,416 undergraduate students were coded as 0; 887 graduate students were coded as 1.


x2(Gender)
x3(Level)
x10(Course is)
x11(Grade)
y2(Q2)
y4(Q4)
x13(Q5)
x14(Q6)
x15(Q7)
x16(Q8)
x17(Q9)
x18(Q12)
x19(Q13)
x20(Q15)

Call #

Gender:
Ⓜ Male
Ⓕ Female

Level:
○ Freshman
○ Sophomore
○ Junior
○ Senior
○ Graduate

Course is:
○Elective ○Required

Expected Grade:
Ⓐ Ⓑ Ⓒ Ⓓ Ⓕ

Code

North Dakota State University
Student Rating of Instruction

Directions:
Using a #2 pencil only, blacken the bubble that best represents your response to each item.

Response Scale for Items 1-6 (from left to right):
VP=Very Poor, P=Poor, IB=In Between, G=Good, VG=Very Good

	VP	P	IB	G	VG
1. Your satisfaction with the instruction in this course	○	○	○	○	○
2. The instructor as a teacher	○	○	○	○	○
3. The ability of the instructor to communicate effectively	○	○	○	○	○
4. The quality of this course	○	○	○	○	○
5. The fairness of procedures for grading this course	○	○	○	○	○
6. Your understanding of the course content	○	○	○	○	○

Response Scale for Items 7-16: SD=Strongly Disagree, D=Disagree, N-Neutral, A=Agree, SA=Strongly Agree

	SD	D	N	A	SA
7. This instructor created an atmosphere that is conducive to learning	○	○	○	○	○
8. This instructor provided well-defined course objectives	○	○	○	○	○
9. This instructor provided content and materials that were clear and well organized	○	○	○	○	○
10. I understood how my grades were assigned in this course	○	○	○	○	○
11. I met or exceeded the course objectives given for this course	○	○	○	○	○
12. The instructor was available to assist students outside of class	○	○	○	○	○
13. The instructor provided feedback in a timely manner	○	○	○	○	○
14. The instructor provided relevant feedback that helped me learn	○	○	○	○	○
15. The instructor set and maintained high standards that students must meet	○	○	○	○	○
16. The physical environment was conductive to learning	○	○	○	○	○

Note for numeric variable: x_{12} represents class size;

Notes for dummy variables:

x_1	represents gender of instructors (0 for female, 1 for male);
x_2	represents gender of students (0 for female, 1 for male);
x_3	represents class level (0 for undergraduate, 1 for graduate);

x_4-x_9	represent colleges: 000000 for College of Agriculture, Food, and Natural Resources Management;
	100000 for College of Arts Humanities and Social Sciences;
	010000 for College of Business;
	001000 for College of Engineering;
	000100 for College of Human Development and Education;
	000010 for College of Health Professions ;
	000001 for College of Science and Mathematics;

x_{10}	represents course type (0 for elective, 1 for required);
x_{21}	represents $x_1 * x_2$

Figure 3.1. Questions for Student Rating of Instruction

The fourth variable was the college that the students attended. There were seven colleges in total: the College of Agriculture, Food, and Natural Resources Management (Ag., Food & NRM) was treated as the baseline and coded as 000000, the College of Arts Humanities and Social Sciences was coded as 100000, the College of Business was coded as 010000, the College of Engineering was coded as 001000, the College of Human Development and Education was coded as 000100, the College of Health Professions was coded as 000010, and the College of Science and Mathematics was coded as 000001.

The fifth variable was the course type. Here, 0 represented elective and 1for required. Of respondents, 22,911 students were taking required courses, and 7,392 students were taking electives. The sixth variable was the student's expected grade for the course. The seventh variable was the size of the class. This was treated as a numerical variable. The size of the class ranged from 12 to 272 students.

3.3. The Development of the Proportion Odds Cumulative Logit Model

3.3.1. Introduction to Proportion Odds Ratio Method

If the response scale for the dependent variable is a set of possible categories, the response is called polytomous. Unlike binary variables, polytomous variables have more than two categories. Usually, there are three types of measurement scales for response variables: (1) nominal scales in which the scale values represent descriptive categories, (2) ordinal scales in which the categories are ordered, and (3) interval scales in which the scale values are ordered with the equal scale unit.

Multinomial logistic regression represents how polytomous dependent variables rely on the independent variables. We need to distinguish whether the response is ordinal or nominal when we analyze a multinomial response. The ordinal scales occur more frequently than other scales. In our study, the response categories for the dependent variable are ordinal scales, which suggests a certain relationship between them. There are five categories for response variable, ranging from “very poor” to “very good” or from “strongly disagree” to “strongly agree.” The proportional-odds cumulative logit model was taken to estimate the effects of different variables on students’ rating for our study. The cumulative response probabilities can be represented as follows:

$$P_i = \sum_{k=1}^{i} p_k$$

$$P_1 = p_1;$$

$$P_2 = p_1 + p_2;$$

$$P_3 = p_1 + p_2 + p_3;$$

$$P_4 = p_1 + p_2 + p_3 + p_4;$$

$$P_5 = p_1 + p_2 + p_3 + p_4 + p_5 = 1;$$

The cumulative logit model can be defined as follows:

$$log(\frac{P_i}{1-P_i}) = \pi_i + x'\beta, i = 1,\ 2,\ 3,\ 4\ ;$$

where $\pi_1 < \pi_2 < \pi_3 < \pi_4$ are intercepts and β is the coefficient vector to be estimated. The β does not depend on the response level.

3.3.2. Model Development

Each student’s rating of the instructor’s performance is considered with the demographic information of instructor and student, and class information for analyzing the effect of genders and gender interaction in our study. The students’ rating for “The instructor as a teacher” and “The quality of this course” were treated individually as dependent variables. Two types of models were set up based on the class information, student’s performance and instructor’s performance for two dependent variables. For the first response variable: (1) the first model type only considered the class information; (2) the second model type not only considered the class information, but also the performance of instructor. For the second response variable: (1) the first model type only considered the class information and student's performance; (2) The second model type includes all the information from the SROI. The relationship between students' gender and instructors' gender was also considered in our models because one of our goals was to illustrate the relationship between these two variables. All related effect of covariates and the main effect of gender were used to analyze the effect of gender.

Backward elimination was used to fit regression models and determine which of the predictor variables had significant effects on the two dependent variables. The backward elimination process kept removing the variable that had the least significant effect until all effects in the model met the specified remaining level.

In our data, 80% of the ratings (24,242 observations) were randomly selected as training data to construct a predictive model. The rest of the data (6,061 observations) was used to test the utility of the model.

4. RESULTS

4.1. Exploratory Analysis

In this study, instructors for 990 courses of seven colleges at NDSU were evaluated by 30,303 students in 2013 and 2014. There were two response variables: y_2 (“The instructor as a teacher”) and y_4 (“The quality of this course”), and all of the analyses focused on these two response variables independently.

We began the research by analyzing the overall average rating by genders of instructors for two response variables. The overall average evaluation score for male instructors was 4.203 for y_2 and 4.078 for y_4, while the average evaluation score for female instructors was 4.161 for y_2 and 4.073 for y_4, as shown in Table 4.1. There was not a significant difference in the overall average ratings by genders for each response variable, but male instructors received slightly higher ratings than female instructors for both y_2 and y_4.

Table 4.1. Ratings by Genders of Instructors for y_2 and y_4

Average Ratings	**Female Instructor**	**Male Instructor**
y_2	4.161	4.203
y_4	4.073	4.078

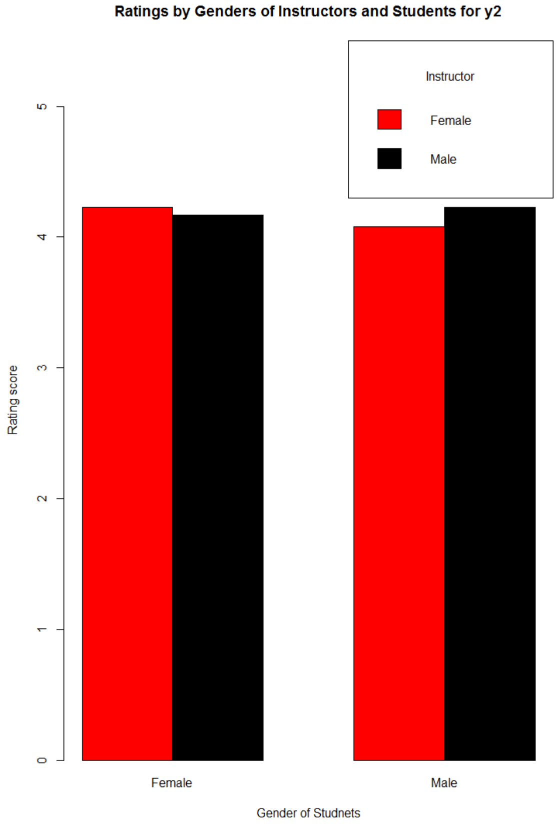
4.1.1. Ratings by Genders of Instructors and Students

We aimed to determine whether there was a relationship between the genders of instructors and students. Based on the results of Pearson’s Chi-Squared test, we found that the p-value was less than 0.05, which indicated a significant correlation between these two variables. We therefore took the interaction between genders of students and instructors into consideration. The comparison between the average rating scores by genders of instructors and students for y_2 and y_4

is shown in Table 4.2, and Figure 4.1 and 4.2. Female students gave higher evaluations to female instructors than male instructors (4.228 for females versus 4.17 for males; 4.162 for females versus 4.067 for males) and male students gave higher evaluations to male instructors (4.229 for males versus 4.079 for females; 4.087 for males versus 3.964 for females) for y_2 and y_4.

Table 4.2. Average Rating by Genders of Instructors and Students for y_2 and y_4

Average Ratings	**Genders**	**Female Instructor**	**Male Instructor**
y_2	Female Student	4.228	4.170
	Male Student	4.079	4.229
y_4	Female Student	4.162	4.067
	Male Student	3.964	4.087



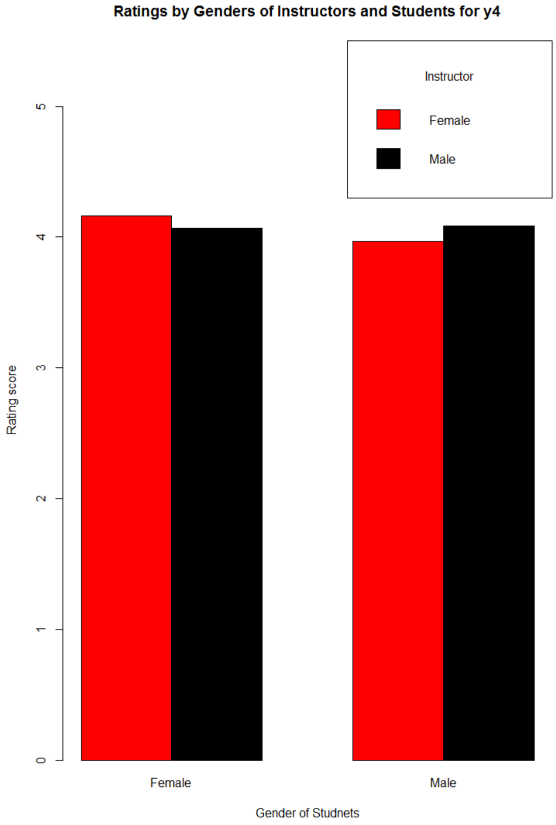


Figure 4.1. Ratings by Genders for y_2

Figure 4.2. Ratings by Genders for y_4

We were also curious about the proportion of each rating category in relation to total ratings. In other words, we had five categories of rating, and the sum of all the proportions should be equal to 1. The proportion of how students rated in each rating category by genders of instructors and students for the first response variable y_2 ("The instructor as a teacher") is shown by four different student-instructor gender combinations in Table 4.3 and Figure 4.3. The trend of each combination is similar. The highest rating score was rated by female students for female instructors (49.93%), which was much higher than the score that female students rated for male instructors (44.99%). However, male students' evaluation for male instructors (47.36% as Very Good) was much higher than for female instructors (42.05% as Very Good). More female students preferred to rate female instructors very highly, whereas more male students preferred to rate male instructors very highly. Less male students rated female instructors highly (42.05% as Very Good) compared to the other three gender combinations: female students rated female instructors (49.93%), female students rated male instructors (44.99%), male students rated male instructors (47.36%).

Table 4.3. Proportions of Student Ratings for y_2 by Genders

Percentage	**1 (Very Poor)**	**2 (Poor)**	**3 (In Between)**	**4 (Good)**	**5 (Very Good)**
StuF-vs-InsF	2.06%	4.84%	11.24%	31.94%	49.93%
StuF-vs-InsM	1.68%	5.01%	12.97%	35.36%	44.99%
StuM-vs-InsF	3.13%	6.22%	12.38%	36.22%	42.05%
StuM-vs-InsM	1.75%	4.07%	11.03%	35.80%	47.36%

A mosaic plot visualizes categorical data in multiple dimensions and illustrates the association between variables, with the relative frequency displayed as rectangular cells. When variables are independent, the rectangles will appear as identical to one another. A mosaic plot was used in this study to visualize how students' ratings were distributed.

Most of the ratings were concentrated in the 4th (Good) and 5th (Very Good) categories for response variable y_2. There were more male students than female students in our data. The rectangles of female students who rated female instructors are bigger than those for female students who rated male instructors in the 4th (Good) and 5th (Very Good) categories, which indicates that more female students gave higher ratings to female instructors. Conversely, male students favored male instructors over female instructors. As the rectangles are not identical, so it indicates genders of students and instructors were associated with each other for y_2.

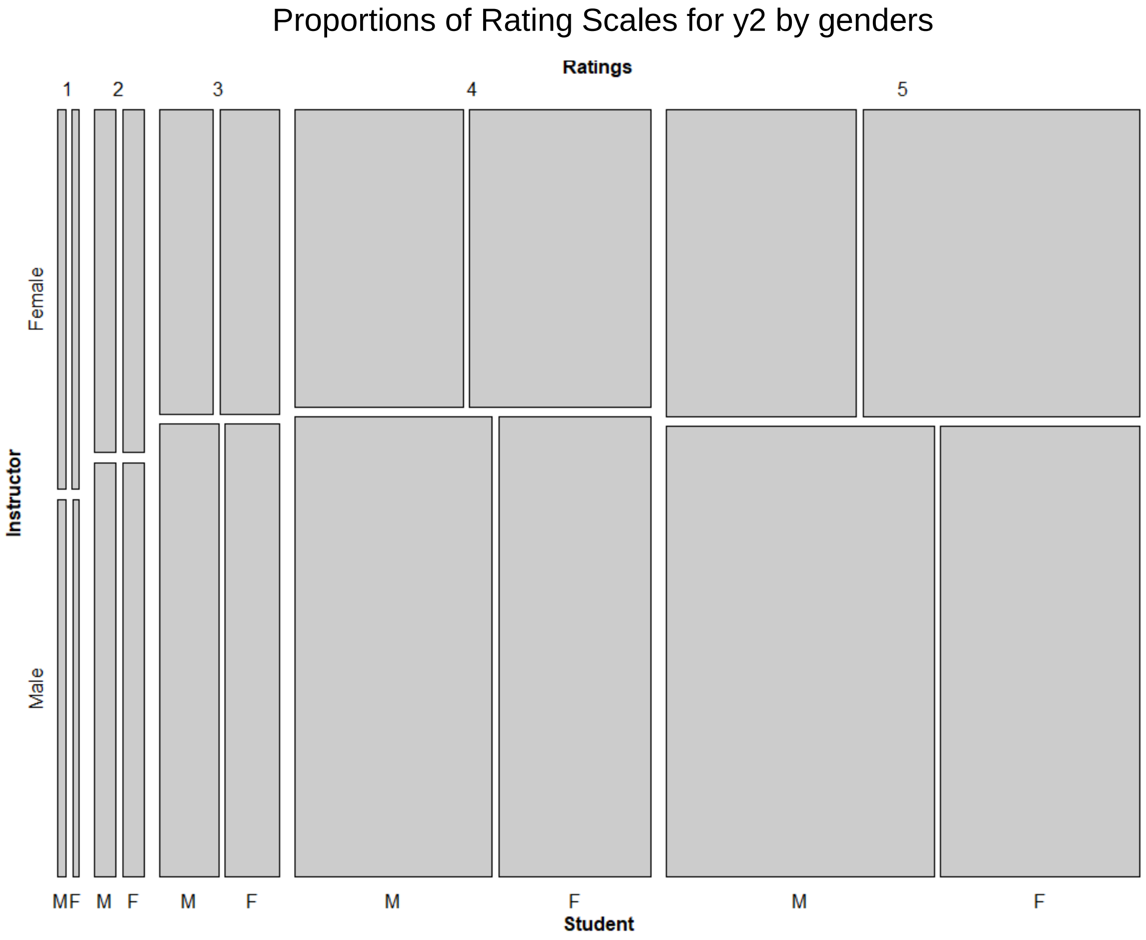


Figure 4.3. Proportions of Student Ratings for y_2 by Genders

The proportion of student ratings in each category for y_4 ("The quality of this course") is compared in Table 4.4 and Figure 4.4. Similar results as the analysis for y_2 were produced, and illustrate that male students favored male instructors (37%) while female students favored female instructors in the 5th rating category (Very Good). More female students rated female instructors in the 5th category (41.3%). The total percentage of the 4th and 5th categories for female students who rated female instructors was approximately 81%, which is only slightly higher than that of female students who rated male instructors (78%) and male students who rated male instructors (79%). However, only 74% of male students rated female instructors in the 4th and 5th categories.

Table 4.4. Proportions of Student Ratings for y_4 by Genders

Percentage	**1** **(Very Poor)**	**2** **(Poor)**	**3** **(In Between)**	**4** **(Good)**	**5** **(Very Good)**
StuF-InsF	1.23%	3.78%	13.83%	39.89%	41.28%
StuF-InsM	1.44%	4.53%	15.94%	42.12%	35.97%
StuM-InsF	2.25%	6.36%	16.75%	42.00%	32.64%
StuM-InsM	1.58%	4.65%	14.72%	41.58%	37.47%

The proportion of rating categories for y_4 in the mosaic plot shows a significant relationship between genders of students and instructors. The proportions of female students who rated female instructors and male students who rated male instructors in the 4th (Good) and 5th (Very Good) categories were greater than the female students who rated male instructors and male students who rated female instructors. The results indicate that the 4th and 5th rating categories for male instructors were mainly from male students, while for female instructors, the highest ratings were given mostly by female students.

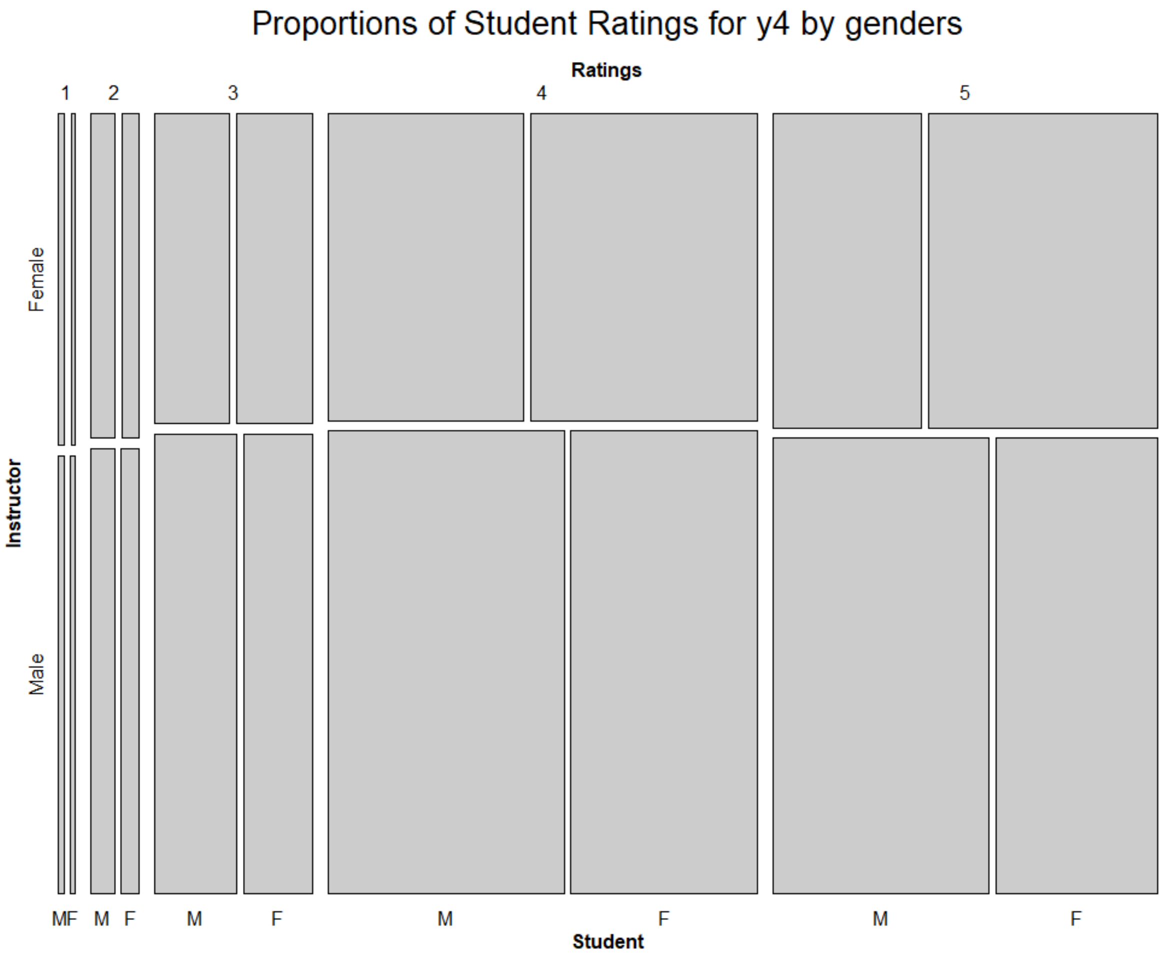


Figure 4.4. Proportions of Student Ratings for y_4 by Genders

4.1.2. Comparison of SROI in Different Colleges

The college attended is an important factor in analyzing potential gender bias in SROI. We compared the average ratings in seven different colleges to determine whether gender biases existed and related to the colleges. The students' distributions in colleges of our study were: (1) 1,669 male students and 1,645 female students from the College of Agriculture, Food, and Natural Resources Management (Ag., Food & NRM). (2) 3,572 male students and 3,569 female students from the College of Arts, Humanities and Social Sciences (AHSS). (3) 1,900 male students and 1,310 female students from the College of Business. (4) 1,859 male students and 416 female students from the College of Engineering. (5) 1,413 male students and 2,046 female students from the College of Human Development and Education (HDE). (6) 326 male students and 735 female students from the College of Health Professions. (7) 4,882 male students and 4,961 female students from the College of Science and Mathematics (Science & Math).

The comparison of average ratings for y_2 ("The instructor as a teacher") between genders of instructors by colleges is shown in Figure 4.5 and Table 4.5. The average ratings of instructors by students in the College of Engineering and the College of Science & Math showed noticeable differences for female and male instructors. In the College of Engineering, female instructors only received an average rating of 2.6, which was the lowest score among all seven colleges, while the average score for male instructors was 4.15. This substantially lower rating for female instructors is clearly illustrated in Figure 4.5. The instructors' distributions in the College of Engineering were 2,166 male instructors and 115 female instructors. Instructors were evaluated by students that 2,259 students as undergraduate level and 16 as graduate level. Among all the students, 367 students were elective to attend the course and 1,908 students were required to attend. The ratings of instructors are similar and high for male (4.424) and female (4.414) instructors in College of Human Development and Education. For response variable y_2, female instructors in the College of Arts, Humanities and Social Sciences, the College of Business, the College of Engineering and the College of Science and Mathematics received lower average ratings compared to male instructors in those colleges.

Table 4.5. Average Rating for y_2 by Colleges

Colleges	Female Instructor	Male Instructor
Ag., Food & NRM	4.272	4.235
AHSS	4.254	4.359
Business	4.084	4.219
Engineering	2.600	4.150
HDE	4.424	4.414
Health Professions	4.346	4.346
Science & Math	3.890	4.074

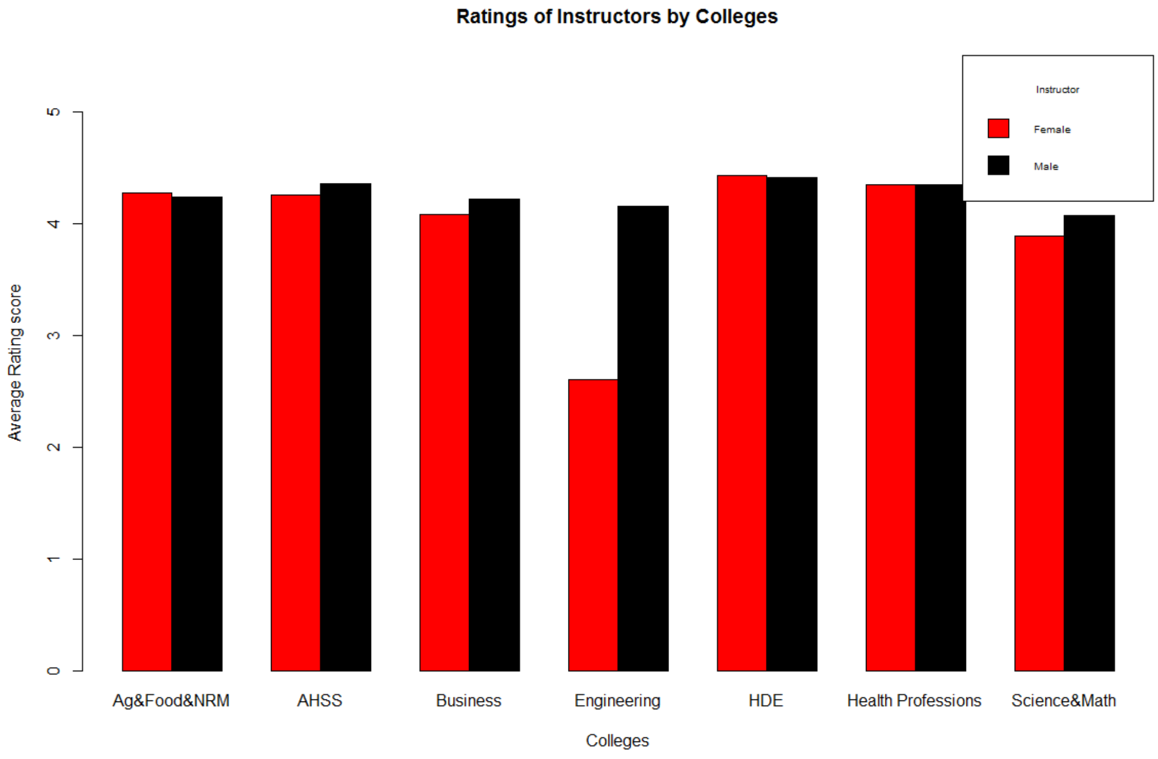


Figure 4.5. Average Rating for y_2 by Colleges

The comparison of average rating for y_4 (“The quality of this course”) between genders of instructors by colleges is shown in Figure 4.6 and Table 4.6. The difference in average ratings for male instructors and female instructors was the most obvious in the College of Engineering, in which female instructors received an average score of only 2.913 for y_4, while male instructors received 4.056 on average. In addition, for response variable y_4, students had high expectations for male and female instructors from the College of Science and Mathematics, as instructors received lower average ratings (3.84, 3.93) when compared with the other colleges (all of which had ratings higher than 4) except the College of Engineering. The averages of the College of Science and Mathematics are obviously lower than the averages of other colleges, excluding that for female instructors at the College of Engineering.

Table 4.6. Average Rating for y_4 by Colleges

Colleges	**Female Instructor**	**Male Instructor**
Ag., Food & NRM	4.195	4.148
AHSS	4.165	4.190
Business	4.004	4.151
Engineering	2.913	4.056
HDE	4.249	4.308
Health Professions	4.296	4.197
Science & Math	3.843	3.927

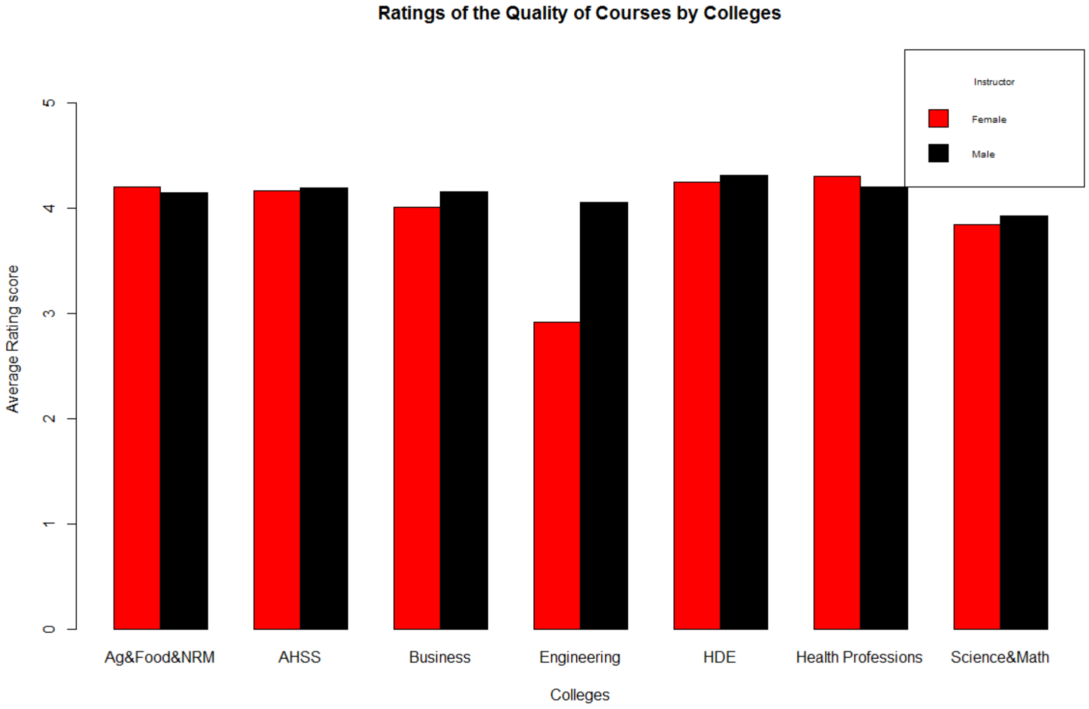


Figure 4.6. Average Rating for y_4 by Colleges

4.2. Modeling Selection

4.2.1. Proportional Odds Ratio Model

Tables 4.7-4.10, below, show the four models we generated using a proportional odds ratio method corresponding to two response variables. In SROI, question 2, how do you rate "The instructor as a teacher", was the first response variable and was represented as y_2. The students' satisfaction for question 4, how to rate "The quality of this course", was the second response variable, y_4. Class information, student's performance and instructor's performance were treated as explanatory variables in the models shown below. The class information included six indicator variables and one numeric variable, as follows: (1) Gender of instructors was coded as 0 for female and 1 for male. (2) Gender of students was coded as 0 for female and 1 for male. (3) Class level was coded as 0 for undergraduate and 1 for graduate. (4) Colleges were coded as follows: The

College of Agriculture, Food, and Natural Resources Management was coded as 000000. The College of Arts Humanities and Social Sciences was coded as 100000. The College of Business was coded as 010000. The College of Engineering was coded as 001000. The College of Human Development and Education was coded as 000100. The College of Health Professions was coded as 000010. The College of Science and Mathematics was coded as 000001. (5) Course type was coded as 0 for elective and 1 for required. (6) Class size was represented by x_1. The performance of the student (expected grade) was represented by x_2.

The performance of instructors was represented by $x_3 - x_{10}$ as follows: (1) The rating for the fairness of procedures for grading this course was represented by x_3. (2) The understanding of the course content was represented by x_4. (3) The evaluation of whether the instructor created an atmosphere that was conducive to learning was represented by x_5. (4) The rating for whether the instructor provided well-defined course objectives was represented by x_6. (5) The rating for whether the instructor provided content and materials that were clear and well organized was represented by x_7. (6) Whether the instructor was available to assist students outside of class was represented by x_8. (8) Whether the instructor provided feedback in a timely manner was represented by x_9. (9) The evaluation of whether the instructor set and maintained high standards that students must meet was represented by x_{10}. The relationship between genders of instructors and students was also taken into consideration.

The intercepts of the proportional odds ratio model can differ, but different equations have the same slope for each variable. For this reason, 'β' represented the constant slope and 'x' represented the effect of independent variables.

The first stage of modeling focused on the effects of variables on the first response variable y_2: "Instructor as a teacher". The optimized predictive model (Model 2) for y_2 had four estimated equations that used the following variables: genders of students and instructors (1 for male, 0 for female), course level (1 for graduate, 0 for undergraduate), class type (1 for required, 0 for elective), class size (x_1), college (000000 for Ag., Food & NRM, 100000 for AHSS, 010000 for Business, 001000 for Engineering, 000100 for HDE, 000010 for Health Professions, 000001 for Science & Math), expected grade (x_2), and the interaction between genders of students and instructors (1 for male, 0 for female), as shown in Table 4.7.

After backward elimination, the effects of coefficients were calculated and are shown in Table 4.8. Most of the covariates had significant effects on our model. The p-value for the gender of instructor was 0.8455, but the interaction between genders of students and instructors was significant.

Table 4.7. Model 2 Includes Class Information for y_2

$$log(\frac{\hat{P}_1}{1-\hat{P}_1}) = log(\frac{\hat{p}_1}{\hat{p}_2+\hat{p}_3+\hat{p}_4+\hat{p}_5}) = -\ 5.370\ +\ x'\hat{\beta}$$

$$log(\frac{\hat{P}_2}{1-\hat{P}_2}) = log(\frac{\hat{p}_2+\hat{p}_1}{\hat{p}_3+\hat{p}_4+\hat{p}_5}) = -\,4.069\ +\ x'\hat{\beta}$$

$$log(\frac{\hat{P}_3}{1-\hat{P}_3}\ ) = log(\frac{\hat{p}_3+\hat{p}_2+\hat{p}_1}{\hat{p}_4+\hat{p}_5}) = -\ 2.856\ +\ x'\hat{\beta}$$

$$log(\frac{\hat{P}_4}{1-\hat{P}_4}\ ) = log(\frac{\hat{p}_4+\ \hat{p}_3+\hat{p}_2+\hat{p}_1}{\hat{p}_5}) = -\ 1.137\ +\ x'\hat{\beta}$$

$$\begin{aligned} x'\hat{\beta} = {} & -0.0071\,\mathbb{1}^{(Instructor)}_{\{Male\}} + 0.2331\,\mathbb{1}^{(Student)}_{\{Male\}} - 0.0592\,\mathbb{1}^{(College)}_{\{\text{AHSS}\}} + 0.2324\,\mathbb{1}^{(College)}_{\{\text{Business}\}} \\ & + 0.5504\,\mathbb{1}^{(College)}_{\{\text{Engineering}\}} - 0.2483\,\mathbb{1}^{(College)}_{\{\text{HDE}\}} - 0.0724\,\mathbb{1}^{(College)}_{\{\text{Health Professions}\}} \\ & + 0.3017\,\mathbb{1}^{(College)}_{\{\text{Science \& Math}\}} - 0.2389\,\mathbb{1}^{(Course)}_{\{Required\}}\ +0.0027\,x_1 + 0.5295\,x_2 \\ & - 0.4544\,\mathbb{1}^{(Instructor)}_{\{Male\}}\mathbb{1}^{(Student)}_{\{Male\}} \end{aligned}$$

Table 4.8. P-Value for Coefficients of Model 2

Covariates	$\mathbb{1}_{\{Male\}}^{(Instructor)}$	$\mathbb{1}_{\{Male\}}^{(Student)}$	$\mathbb{1}_{\{Required\}}^{(Course)}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$
P-Value	0.8455	<0.0001	<0.0001	<0.0001	<0.0001
Covariates	$\mathbb{1}_{\{\text{AHSS}\}}^{(College)}$	$\mathbb{1}_{\{\text{Business}\}}^{(College)}$	$\mathbb{1}_{\{\text{Engineering}\}}^{(College)}$	$\mathbb{1}_{\{\text{HDE}\}}^{(College)}$	
P-Value	0.1951	<0.0001	<0.0001	<0.0001	
Covariates	$\mathbb{1}_{\{\text{Science \& Math}\}}^{(College)}$		$\mathbb{1}_{\{\text{Health Professions}\}}^{(College)}$	$\mathbb{1}_{\{Male\}}^{(Instructor)}\mathbb{1}_{\{Male\}}^{(Student)}$	
P-Value	<0.0001		0.3441	<0.0001	

The ratings for the performance of instructors were then included in our model. The optimized predictive model (Model 4) for y_2 is shown in Table 4.9, and used the following variables: genders of students and instructors (1 for male, 0 for female), class type (1 for required, 0 for elective), class size (x_1), college (000000 for Ag., Food & NRM, 100000 for AHSS, 010000 for Business, 001000 for Engineering, 000100 for HDE, 000010 for Health Professions, 000001 for Science & Math), variables for the performance of instructors (from x_3 through x_{10}), expected grade (x_2), and the interaction between genders of students and instructors (1 for male, 0 for female).

After backward elimination, the effects of coefficients were calculated and are shown in Table 4.10. Most of the covariates had significant effects on our model. Most of the coefficients for instructor's performance (from x_3 through x_{10}) were less than 0.0001.

Table 4.9. Model 4 Involves Instructor’s Performance from SROI for y_2

$log(\frac{\hat{P}_1}{1-\hat{P}_1}) = log(\frac{\hat{p}_1}{\hat{p}_2+\hat{p}_3+\hat{p}_4+\hat{p}_5}) = 8.6151 + x'\hat{\beta}$
$log(\frac{\hat{P}_2}{1-\hat{P}_2}) = log(\frac{\hat{p}_2+\hat{p}_1}{\hat{p}_3+\hat{p}_4+\hat{p}_5}) = 11.2823 + x'\hat{\beta}$
$log(\frac{\hat{P}_3}{1-\hat{P}_3}) = log(\frac{\hat{p}_3+\hat{p}_2+\hat{p}_1}{\hat{p}_4+\hat{p}_5}) = 13.8427 + x'\hat{\beta}$
$log(\frac{\hat{P}_4}{1-\hat{P}_4}) = log(\frac{\hat{p}_4+\hat{p}_3+\hat{p}_2+\hat{p}_1}{\hat{p}_5}) = 17.2288 + x'\hat{\beta}$
$x'\hat{\beta} = -0.1083\, \mathbb{1}_{\{Male\}}^{(Instructor)} + 0.1789\, \mathbb{1}_{\{Male\}}^{(Student)} - 0.1097\, \mathbb{1}_{\{\text{AHSS}\}}^{(College)} + 0.1415\, \mathbb{1}_{\{\text{Business}\}}^{(College)}$ $+ 0.2901\, \mathbb{1}_{\{\text{Engineering}\}}^{(College)} - 0.2304\, \mathbb{1}_{\{\text{HDE}\}}^{(College)} - 0.1542\, \mathbb{1}_{\{\text{Health Professions}\}}^{(College)}$ $+ 0.1151\, \mathbb{1}_{\{\text{Science \& Math}\}}^{(College)} - 0.1269\, \mathbb{1}_{\{Required\}}^{(Course)} + 0.0004\, x_1 - 0.1378\, x_2$ $- 0.4302\, x_3 - 0.3803\, x_4 - 1.3171\, x_5 - 0.2395\, x_6 - 0.7451\, x_7 - 0.2057\, x_8$ $- 0.1851\, x_9 - 0.4451\, x_{10} - 0.3384\, \mathbb{1}_{\{Male\}}^{(Instructor)} \mathbb{1}_{\{Male\}}^{(Student)}$

Table 4.10. P-Value for Coefficients of Model 4

Covariates	$\mathbb{1}_{\{Male\}}^{(Instructor)}$	$\mathbb{1}_{\{Male\}}^{(Student)}$	$\mathbb{1}_{\{Required\}}^{(Course)}$	x_1	x_2	
P-Value	0.0150	0.0001	< 0.0001	0.1446	< 0.0001	
Covariates	$\mathbb{1}_{\{\text{AHSS}\}}^{(College)}$	$\mathbb{1}_{\{\text{Business}\}}^{(College)}$	$\mathbb{1}_{\{\text{Engineering}\}}^{(College)}$	$\mathbb{1}_{\{\text{HDE}\}}^{(College)}$	$\mathbb{1}_{\{\text{Health Professions}\}}^{(College)}$	
P-Value	0.0506	0.0269	<0.0001	0.0007	0.1045	
Covariates	x_3	x_4	x_5	x_6	x_7	x_8
P-Value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Covariates	x_9	x_{10}	$\mathbb{1}_{\{Male\}}^{(Instructor)} \mathbb{1}_{\{Male\}}^{(Student)}$		$\mathbb{1}_{\{\text{Science \& Math}\}}^{(College)}$	
P-Value	<0.0001	<0.0001	< 0.0001		0.0303	

The second stage of modeling focused on the effects of variables on the second dependent variable y_4 ("The quality of this course"). Initially, we only considered the effects of explanatory variables, which were the genders of students and instructors (1 for male, 0 for female), class type (1 for required, 0 for elective), class size (x_1), college (000000 for Ag., Food & NRM, 100000 for AHSS, 010000 for Business, 001000 for Engineering, 000100 for HDE, 000010 for Health Professions, 000001 for Science & Math), expected grade (x_2), and the interaction between genders of students and instructors (1 for male, 0 for female). The first optimized model (Model 6) for y_4 is shown in Table 4.11. The effects of coefficients were calculated after backward elimination, and are shown in Table 4.12. Most of the covariates had significant effects on our model. Variable for gender of instructor remained in the model because the interaction between genders of instructors and students had a significant effect on the model.

Table 4.11. Model 6 Includes Class Information and Student's Performance for y_4

$log(\frac{\hat{P}_1}{1-\hat{P}_1}) = log(\frac{\hat{p}_1}{\hat{p}_2+\hat{p}_3+\hat{p}_4+\hat{p}_5}) = -\ 5.9177\ +\ x'\hat{\beta}$
$log(\frac{\hat{P}_2}{1-\hat{P}_2}) = log(\frac{\hat{p}_2+\hat{p}_1}{\hat{p}_3+\hat{p}_4+\hat{p}_5}) = -\,4.4266\ +\ x'\hat{\beta}$
$log(\frac{\hat{P}_3}{1-\hat{P}_3}\ ) = log(\frac{\hat{p}_3+\hat{p}_2+\hat{p}_1}{\hat{p}_4+\hat{p}_5}) = -\,2.9377\ +\ x'\hat{\beta}$
$log(\frac{\hat{P}_4}{1-\hat{P}_4}\ ) = log(\frac{\hat{p}_4+\ \hat{p}_3+\hat{p}_2+\hat{p}_1}{\hat{p}_5}) = -\ 0.9693\ +\ x'\hat{\beta}$
$x'\hat{\beta} = 0.0692\ \mathbb{1}_{\{Male\}}^{(Instructor)} + 0.3035\ \mathbb{1}_{\{Male\}}^{(Student)} +\ 0.0342\ \ \mathbb{1}_{\{\text{AHSS}\}}^{(College)} + 0.1544\ \mathbb{1}_{\{\text{Business}\}}^{(College)}$ $+\ 0.5216\ \mathbb{1}_{\{\text{Engineering}\}}^{(College)} - 0.0483\ \mathbb{1}_{\{\text{HDE}\}}^{(College)} - 0.1008\ \mathbb{1}_{\{\text{Health Professions}\}}^{(College)}$ $+\ 0.4317\ \mathbb{1}_{\{\text{Health Professions}\}}^{(College)} - 0.2456\ \mathbb{1}_{\{Required\}}^{(Course)} + 0.0015\ x_1 + 0.6211\ x_2$ $-\ 0.4255\ \mathbb{1}_{\{Male\}}^{(Instructor)}\mathbb{1}_{\{Male\}}^{(Student)}$

Table 4.12. P-Value for Coefficients of Model 6

Covariates	$\mathbb{1}_{\{Male\}}^{(Instructor)}$	$\mathbb{1}_{\{Male\}}^{(Student)}$	$\mathbb{1}_{\{Required\}}^{(Course)}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$
P-Value	0.0531	<0.0001	<0.0001	<0.0001	<0.0001
Covariates	$\mathbb{1}_{\{\mathrm{AHSS}\}}^{(College)}$	$\mathbb{1}_{\{\mathrm{Business}\}}^{(College)}$	$\mathbb{1}_{\{\mathrm{Engineering}\}}^{(College)}$	$\mathbb{1}_{\{\mathrm{HDE}\}}^{(College)}$	
P-Value	0.4457	0.0029	<0.0001	0.3653	
Covariates	$\mathbb{1}_{\{\mathrm{Science\ \&\ Math}\}}^{(College)}$		$\mathbb{1}_{\{\mathrm{Health\ Professions}\}}^{(College)}$	$\mathbb{1}_{\{Male\}}^{(Instructor)}\mathbb{1}_{\{Male\}}^{(Student)}$	
P-Value	<0.0001		0.1892	<0.0001	

The second optimized model (Model 7) used all variables, including genders of students and instructors (1 for male, 0 for female), course level (1 for graduate, 0 for undergraduate), class type (1 for required, 0 for elective), class size (x_1), college (000000 for Ag., Food & NRM, 100000 for AHSS, 010000 for Business, 001000 for Engineering, 000100 for HDE, 000010 for Health Professions, 000001 for Science & Math), variables for the performance of instructors (from x_3 through x_{10}), expected grade (x_2), and the interaction between genders of students and instructors (1 for male, 0 for female). Backward elimination determined that all of the variables had significant effects on y_4, so all variables remained in the optimized model (Model 7) shown in Table 4.13. The effects of coefficients were calculated and are shown in Table 4.14. The effect of whether the instructor provided feedback in a timely manner (x_9) was 0.0634, and it remained in the model. The reason is that backward elimination used the Akaike information criterion (AIC), and the AIC maintained x_9 as an important covariate of the model. Most of the covariates had significant effects on our model.

Table 4.13. Model 7 Includes All of the Information from SROI for y_4

$log(\frac{\hat{P}_1}{1-\hat{P}_1}) = log(\frac{\hat{p}_1}{\hat{p}_2+\hat{p}_3+\hat{p}_4+\hat{p}_5}) = 8.1062 + x'\hat{\beta}$
$log(\frac{\hat{P}_2}{1-\hat{P}_2}) = log(\frac{\hat{p}_2+\hat{p}_1}{\hat{p}_3+\hat{p}_4+\hat{p}_5}) = 10.7958 + x'\hat{\beta}$
$log(\frac{\hat{P}_3}{1-\hat{P}_3}) = log(\frac{\hat{p}_3+\hat{p}_2+\hat{p}_1}{\hat{p}_4+\hat{p}_5}) = 13.6963 + x'\hat{\beta}$
$log(\frac{\hat{P}_4}{1-\hat{P}_4}) = log(\frac{\hat{p}_4+\hat{p}_3+\hat{p}_2+\hat{p}_1}{\hat{p}_5}) = 17.3927 + x'\hat{\beta}$
$x'\hat{\beta} = 0.0347\mathbb{1}_{\{Male\}}^{(Instructor)} + 0.3245\,\mathbb{1}_{\{Male\}}^{(Student)} + 0.1497\,\mathbb{1}_{\{Graduate\}}^{(Course)} + 0.1688\,\mathbb{1}_{\{\text{AHSS}\}}^{(College)}$ $+ 0.0397\,\mathbb{1}_{\{\text{Business}\}}^{(College)} + 0.2309\,\mathbb{1}_{\{\text{Engineering}\}}^{(College)} + 0.2560\,\mathbb{1}_{\{\text{HDE}\}}^{(College)}$ $- 0.0390\,\mathbb{1}_{\{\text{Health Professions}\}}^{(College)} + 0.2883\,\mathbb{1}_{\{\text{Science \& Math}\}}^{(College)} + 0.1358\,\mathbb{1}_{\{Required\}}^{(Course)}$ $- 0.0008x_1 - 0.1821x_2 - 0.6605\,x_3 - 0.9581\,x_4 - 0.7157\,x_5 - 0.2613\,x_6$ $- 0.4656\,x_7 - 0.1014\,x_8 - 0.0427x_9 - 0.6780x_{10}$

Table 4.14. P-Value for Coefficients of Model 7

Covariates	$\mathbb{1}_{\{Male\}}^{(Instructor)}$	$\mathbb{1}_{\{Male\}}^{(Student)}$	$\mathbb{1}_{\{Graduate\}}^{(Course)}$	$\mathbb{1}_{\{Required\}}^{(Course)}$	x_1	x_2
P-Value	0.4157	<0.0001	0.1055	0.0001	0.0048	<0.0001
Covariates	$\mathbb{1}_{\{\text{AHSS}\}}^{(College)}$	$\mathbb{1}_{\{\text{Business}\}}^{(College)}$	$\mathbb{1}_{\{\text{Engineering}\}}^{(College)}$	$\mathbb{1}_{\{\text{HDE}\}}^{(College)}$	$\mathbb{1}_{\{\text{Health Professions}\}}^{(College)}$	
P-Value	0.0016	0.5191	0.0008	0.0001	0.6856	
Covariates	x_3	x_4	x_5	x_6	x_7	x_8
P-Value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Covariates	x_9	x_{10}	$\mathbb{1}_{\{\text{Science \& Math}\}}^{(College)}$		$\mathbb{1}_{\{Male\}}^{(Instructor)}\mathbb{1}_{\{Male\}}^{(Student)}$	
P-Value	0.0634	<0.0001	<0.0001		0.0001	

4.2.2. Goodness of Fit

We aimed to identify the most well-fitting models for two response variables by further testing the goodness of fit for our models in two different ways: accuracy and dispersion parameter. Accuracy was an important consideration when choosing predictive models in our study; a well-fitting model should have high accuracy. We randomly selected 80% of the ratings (24,242 observations) to train data and construct our predictive models. After backward elimination, we selected the best subset of predictors for four optimized models: Model 2 and Model 4 for y_2, and Model 6 and Model 7 for y_4. The remaining 20% of the ratings (6,061 observations) were used to test the accuracy of the models. We employed our predictive models using all of the information for each student to predict how this student would rate the instructor for the corresponding response variable (y_2 or y_4). We calculated the average ratings in class and then the average ratings of all the classes for two response variables. We compared our predicted ratings with actual ratings, as presented in Table 4.15. All predictions by our models were similar to the actual corresponding rating (4.186, 4.181 compared to 4.189, and 4.075, 4.080 compared to 4.082).

Table 4.15. The Accuracy of Different Models

Actual Average Rating	**Predicted Average Rating**	
4.189	Model 2	4.186
	Model 4	4.181
4.082	Model 6	4.075
	Model 7	4.080

In our study, the dispersion problem was another important concern when selecting effective predictive models, as the dispersion problem distorts overall goodness of fit. The dispersion parameter of a good model should be around 1. All ratings were divided into several groups according to how many character variables were in that model. For Model 2, Model 4 and

Model 6, the variables of genders of instructors and students, course type, college, expected grade and class size were used to restructure testing data for prediction. Model 7 also used the effect of variable class level. Only groups with more than one record should be chosen. There were 231 groups of data remained from which to make predictions.

The predicted probabilities of each of the five rating categories were compared with our models with actual values using the Pearson Chi-Square method. The dispersion parameters of all four models were around 1, which meant that there was neither an over-dispersion nor an under-dispersion problem in these two models, as shown in Table 4.16. According to these two tests' results (accuracy and Pearson Chi-Square test), Model 2 and Model 4 for response variable y_2, and Model 6 and Model 7 for response variable y_4 fit well and became our final optimized models, with high accuracy and no over-dispersion problem.

Table 4.16. Pearson Chi-Square Test for Dispersion Parameter

Model	**Pearson Chi-Square ($\mathcal{X}^2$)**	**DF**	**Dispersion Parameter**
Model 2	942.88	900	1.04
Model 4	944.99	900	1.05
Model 6	982.6	900	1.09
Model 7	1951.71	1992	0.98

4.2.3. Check the Accuracy of Models

To test how accurately the models could predict new data points, we randomly selected ratings in three different classes from different colleges in testing data. The model which could be treated as a benchmark for measuring an instructor's teaching performance should only include the information of the class, instead of including the information of the instructor's and student's performance, because we were not able to obtain instructors' performance at the beginning of each semester. Therefore, Model 2 and Model 4 were selected to make the predictions for these three classes. One class was taught by a female instructor from the College of Arts, Humanities and Social Sciences. It had 37 completed SROI, composed of 21 ratings from female students and 16 ratings from male students. Another class was taught by a male instructor from the College of Science and Mathematics. It had 37 completed SROI, composed of 16 ratings from female students and 21 ratings from male students. The third class was taught by a female instructor from the College of Human Development and Education. It had 40 useful SROI, composed of 13 ratings from female students and 27 ratings from male students.

We predicted how each student would rate the teacher for that class by using our two optimized predictive models (Model 2 and Model 4), which corresponded to two response variables y_2 ("The instructor as a teacher") and y_4 ("The quality of this course"). The final accuracy was calculated by comparing the count of correct predictions with the count of actual ratings of the corresponding response variable in that class. Thus, each model had one accuracy comparison in each class. The usual method for analyzing ordinal ratings of instructions is to calculate the average of ratings, but this does not display how each category changes. The advantage of a proportional odds ratio model is that all categories of ordinal ratings are tested in parallel. We analyzed not only the average response for evaluating instructions, but also the response of each ordinal rating category.

The predicted and actual average score in each category of ratings, and the overall average rating score / predicted rating score for y_2 and y_4 of the class from the College of Arts, Humanities and Social Sciences using Model 2 and Model 4, are shown in Table 4.17. The predicted average rating of Model 2 in each category was lower than the actual rating. There was a difference of 0.5

between our predicted overall rating (4.047) and the actual overall rating (4.568) for y_2. The predicted average rating of Model 4 was similar to the actual rating. There was a difference of 0.041 between our predicted overall rating (4.283) and actual overall rating (4.324) for y_4.

Table 4.17. Predicted Versus Actual Rating for y_2 and y_4 in the Class from the College of Arts, Humanities and Social Sciences

Response Variable	**Model**	**1 (Very Poor)**	**2 (Poor)**	**3 (In Between)**	**4 (Good)**	**5 (Very Good)**	**Overall Average Predicted/ Actual score**
y_2	Model 2	0.025	0.119	0.440	1.529	1.935	4.047
	Actual Rating	0.000	0.000	0.324	0.865	3.378	4.568
y_4	Model 4	0.008	0.053	0.314	1.587	2.321	4.283
	Actual Rating	0.000	0.054	0.324	1.514	2.432	4.324

The comparisons of the count of predicted ratings and actual ratings for Model 2 and Model 4 are shown in Table 4.18 and Table 4.19. The counts on the diagonal mean were predicted correctly; otherwise, they were misclassified. We predicted 17 ratings correctly and 20 ratings incorrectly for y_2. The accuracy for predicting how students would rate the instructor in the class from the College of Arts, Humanities and Social Sciences in predictive Model 2 was 45.9%, as shown in Table 4.18. We predicted 19 ratings of Model 4 incorrectly and 18 ratings correctly. The accuracy for Model 4 to predict the instructor in the class from the College of Agriculture, Food, and Natural Resources Management was 51.4%, as shown in Table 4.19.

Table 4.18. Predicted Versus Actual Rating Counts for y_2 in the Class from the College of Arts, Humanities and Social Sciences

Prediction/ Actual Count	**1 (Very Poor)**	**2 (Poor)**	**3 (In Between)**	**4 (Good)**	**5 (Very Good)**	**Predicted Total Count**
1 (Very Poor)	0	0	0	0	0	0
2 (Poor)	0	0	0	0	0	0
3 (In Between)	0	0	0	0	0	0
4 (Good)	0	0	2	5	13	20
5 (Very Good)	0	0	2	3	12	17
Actual Total Count	0	0	4	8	25	37

Table 4.19. Predicted Versus Actual Rating Counts for y_4 in the Class from the College of Arts, Humanities and Social Sciences

Prediction/ Actual Count	**1 (Very Poor)**	**2 (Poor)**	**3 (In Between)**	**4 (Good)**	**5 (Very Good)**	**Predicted Total Count**
1 (Very Poor)	0	0	0	0	0	0
2 (Poor)	0	0	0	0	0	0
3 (In Between)	0	0	0	0	0	0
4 (Good)	0	1	0	6	5	12
5 (Very Good)	0	0	4	8	13	25
Actual Total Count	0	1	4	14	18	37

The predicted and actual average score in each category of ratings, and the overall average rating score / predicted rating score for y_2 and y_4 in the class from the College of Science and Mathematics using Model 2 and Model 4, are shown in Table 4.20. The predicted average rating of Model 2 and Model 4 in each category was similar to the actual rating. Most of the students rated instructors as Good and Very Good. There was a difference of 0.188 between our predicted overall rating (3.947) and the actual overall rating (4.135) for y_2, and a difference of 0.084 between our predicted overall rating (4.240) and actual overall rating (4.324) for y_4.

Table 4.20. Predicted Versus Actual Rating for y_2 and y_4 in the Class from the College of Science and Mathematics

Response Variable	**Model**	**1 (Very Poor)**	**2 (Poor)**	**3 (In Between)**	**4 (Good)**	**5 (Very Good)**	**Overall Average Predicted/ Actual score**
y_2	Model 2	0.030	0.142	0.500	1.540	1.735	3.947
	Actual Rating	0.027	0.054	0.568	1.189	2.297	4.135
y_4	Model 4	0.009	0.060	0.345	1.620	2.207	4.240
	Actual Rating	0.000	0.000	0.486	1.405	2.432	4.324

The comparisons of the count of predicted ratings and actual ratings for Model 2 and Model 4 are shown in Table 4.21 and Table 4.22. We predicted 16 ratings correctly using Model 2 for y_2. The accuracy for predicting how students would rate this instructor in predictive Model 2 was 43.2%, as shown in Table 4.21. We predicted 18 ratings correctly in Model 4 for y_4. The accuracy for Model 4 to predict the class from the College of Science and Mathematics was 54.1%, as shown in Table 4.22.

Table 4.21. Predicted Versus Actual Rating Counts for y_2 in the Class from the College of Science and Mathematics

Prediction/ Actual Count	**1 (Very Poor)**	**2 (Poor)**	**3 (In Between)**	**4 (Good)**	**5 (Very Good)**	**Predicted Total Count**
1 (Very Poor)	0	0	0	0	0	0
2 (Poor)	0	0	0	0	0	0
3 (In Between)	0	0	0	0	0	0
4 (Good)	1	1	6	8	9	25
5 (Very Good)	0	0	1	3	8	12
Actual Total Count	1	1	7	11	17	37

Table 4.22. Predicted Versus Actual Rating Counts for y_4 in the Class from the College of Science and Mathematics

Prediction/ Actual Count	**1 (Very Poor)**	**2 (Poor)**	**3 (In Between)**	**4 (Good)**	**5 (Very Good)**	**Predicted Total Count**
1 (Very Poor)	0	0	0	0	0	0
2 (Poor)	0	0	0	0	0	0
3 (In Between)	0	0	0	0	0	0
4 (Good)	0	0	3	7	5	15
5 (Very Good)	0	0	3	6	13	22
Actual Total Count	0	0	6	13	18	37

The predicted and actual average rating in each category of ratings, and the overall average rating score / predicted rating score for y_2 and y_4 in the class from the College of Human Development and Education, using Model 2 and Model 4, are shown in Table 4.23. The predicted average rating of Model 2 and Model 4 in each category was close to the actual rating. There was a difference of 0.108 between our predicted overall rating (4.217) and the actual overall rating (4.325) for y_2, and a difference of 0.173 between our predicted overall rating (3.727) and actual overall rating (3.9) for y_4. Our predictions were slightly lower than the actual rating scores for y_2 and y_4.

Table 4.23. Predicted Versus Actual Rating for y_2 and y_4 in the Class from the College of Human Development and Education

Response Variable	**Model**	**1 (Very Poor)**	**2 (Poor)**	**3 (In Between)**	**4 (Good)**	**5 (Very Good)**	**Overall Average Predicted/ Actual score**
y_2	Model 2	0.017	0.086	0.344	1.426	2.344	4.217
	Actual Rating	0.000	0.050	0.525	1.000	2.750	4.325
y_4	Model 4	0.026	0.161	0.723	1.792	1.026	3.727
	Actual Rating	0.000	0.250	0.525	1.500	1.625	3.900

The comparisons of the count of predicted ratings and actual ratings for Model 2 and Model 4 are shown in Table 4.24 and Table 4.25. We predicted 22 ratings correctly when using Model 2 for y_2. The accuracy for predicting how students would rate the instructor in predictive Model 2 was 55%. We predicted 16 ratings correctly in Model 4 for y_4, as shown in Table 4.24. The accuracy for Model 4 to predict the class from the College of Human Development and Education was 40%, as shown in Table 4.25.

Table 4.24. Predicted Versus Actual Rating Counts for y_2 in the Class from the College of Human Development and Education

Prediction/ Actual Count	**1 (Very Poor)**	**2 (Poor)**	**3 (In Between)**	**4 (Good)**	**5 (Very Good)**	**Predicted Total Count**
1 (Very Poor)	0	0	0	0	0	0
2 (Poor)	0	0	0	0	0	0
3 (In Between)	0	0	0	0	0	0
4 (Good)	0	0	3	2	2	7
5 (Very Good)	0	1	4	8	20	33
Actual Total Count	0	1	7	10	22	40

Table 4.25. Predicted Versus Actual Rating Counts for y_4 in the Class from the College of Human Development and Education

Prediction/ Actual Count	**1 (Very Poor)**	**2 (Poor)**	**3 (In Between)**	**4 (Good)**	**5 (Very Good)**	**Predicted Total Count**
1 (Very Poor)	0	0	0	0	0	0
2 (Poor)	0	0	0	0	0	0
3 (In Between)	0	1	1	0	0	2
4 (Good)	0	4	6	15	13	38
5 (Very Good)	0	0	0	0	0	0
Actual Total Count	0	5	7	15	13	40

5. CONCLUSIONS AND DISCUSSION

The purpose of this study is to determine whether gender, class level (graduate or undergraduate), course type (required or elective), college, expected grade (student's performance), class size and instructors' performance influence students' ratings, and to construct a statistical model to precisely determine the way in which these variables affect SROI.

After testing the goodness of fit and accuracy of our models, four optimized models (Model 2, Model 4, Model 6 and Model 7) were developed to analyze the satisfaction of students related to instructor's performance. Model 2 and Model 6 were selected for the use of university administrators, to allow them to evaluate the quality of courses and the ability of instructors.

Model 2 and Model 6 used all the information except the instructor's performance. The Pearson Chi-Square test and mosaic plot proved that genders of students and instructors are associated. Our random selection of three classes from different colleges as examples to check the accuracy of our optimized models determined that Model 2 and Model 4 had predicted scores that were close to the actual ratings (Table 4.17- 4.25).

The models were established using a proportional odds ratio method. The advantage of this method is that it analyzes how each category of ratings is changed, rather than simply measuring average ratings. Furthermore, it can accommodate both response variables and explanatory variables that have multiple categories. There were five categories of the response scale in this study, of which only the first four were used to establish a proportional odds ratio model, because the total of all the probabilities equals 1. The slope of each variable gives the trend of the first four categories' probabilities, so the fifth probability of the response scale has an opposite trend to the other four probabilities, which means that the probability of the highest category for ratings will increase if the other four probabilities decrease. The log odds of a student giving a higher rating are greater when a covariate has a higher coefficient and other covariates remain constant.

The summary of Model 2 for response variable y_2 (“The instructor as a teacher”) demonstrates that college, class type, expected grade, genders of students and instructors, and class size play a significant role (Table 4.9). The coefficient of class size is positive, which means that the instructor will receive more ratings in the first four categories than in the fifth category if the class contains more students, given that all of the other variables in the model remain constant. The coefficient of expected grade is positive, so instructors will receive fewer ratings in the fifth category (Very Good) if the students have higher expectations.

For the second response variable y_4 (“The quality of this course”), from the summary of Model 6, all of the variables help to establish the model (Table 4.13). The coefficient for students from the College of Engineering is the largest compared to other colleges, which means that the instructor will receive more ratings in the first four categories when the students are from the College of Engineering, while other variables remain constant. The coefficients of class size and expected grade are positive, which means that the instructors will receive lower ratings when they are in a large class or when students have higher expectations, when all variables remain constant in the model.

We were interested in how student satisfaction differed in the different genders of students and instructors. The results demonstrated that when the ratings came from male students, the difference between the coefficients for male instructor and female instructor in Model 2 and Model 4 was negative 0.4615 and negative 0.4466, while other covariates remained constant. This means that male students would give fewer ratings in the first four rating categories to male instructors for y_2 (“The instructor as a teacher”).

The coefficients' difference between male students who rated male instructors and male students who rated female instructors in Model 6 and Model 7 was negative 0.122 and negative 0.1911, while other covariates remained constant. This means that male students favored male instructors for the second response variable y_4 ("The quality of this course").

When the instructors' performance was rated by female students, the difference between coefficients for male instructors and female instructors in Model 2 and Model 4 was negative 0.0071 and negative 0.1083, while other covariates remained constant. The results demonstrated that female students had higher expectations of female instructors, as they gave fewer fifth category ratings to female instructors for the first response variable y_2 ("The instructor as a teacher").

The coefficients' difference between female students who rated male and female instructors in Model 6 and Model 7 was positive 0.0692 and positive 0.0347, while other covariates remained constant. This indicates that female students would give more of the first four rating categories to male instructors compared to female instructors for y_4 ("The quality of this course"). In other words, female students favor female instructors for the quality of the course.

Therefore, we cannot easily conclude whether gender bias exists against instructors at NDSU. The final conclusion depends on the composition of different genders of students in a class.

REFERENCES

Anne Boring, Kellie Ottoboni, & Philip B. Stark. (2017). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, *2016*(1), 1–11.

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., & Silberg, N. T. (1988). Student Evaluations of College Professors: Are Female and Male Professors Rated Differently? *Journal of Educational Psychology*, *79*(3), 308–14.

Bennett, S. K. (1982). Student Perceptions of and Expectations for Male and Female Instructors: Evidence Relating to the Question of Gender Bias in Teaching Evaluation. *Journal of Educational Psychology*, *74*(2), 170–79.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, *145*, 27–41.

Bowles, L. T. (2000). The evaluation of teaching. *Medical Teacher*, *22*(3), 221–224.

Centra, J. A., & Gaubatz, N. B. (2000). Is There Gender Bias in Student Evaluations of Teaching? *The Journal of Higher Education*, *71*(1), 17–33.

Dodeen, H. (2013). Validity, Reliability, and Potential Bias of Short Forms of Students' Evaluation of Teaching: The Case of UAE University. *Educational Assessment*, *18*(4), 235–250.

Feit, C. (2014). *Student ratings of instruction and student motivation: Is there a connection?* ProQuest Dissertations Publishing. Retrieved September 30, 2018, from

Griffin, B. W. (2006). Grading Leniency, Grade Discrepancy, and Student Ratings of Instruction. *Contemporary Educational Psychology*, *29*(4), 410–425.

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, *4*(1).

Huebner, L., & Magel, R. C. (2015). A Gendered Study of Student Ratings of Instruction. *Open Journal of Statistics*, *05*(06), 552–567.

Ibrahim, A. M. (2011). Using Generalizability Theory to Estimate the Relative Effect of Class Size and Number of Items on the Dependability of Student Ratings of Instruction. *Psychological Reports*, *109*(1), 252–258.

Jones, F. (2017). Comparing Student, Instructor, Classroom and Institutional Data to Evaluate a Seven-Year Department-Wide Science Education Initiative. *Assessment & Evaluation in Higher Education*, *43*(2), 323–338.

Long, C. S., Ibrahim, Z., & Kowang, T. O. (2015). An Analysis on the Relationship between Lecturers' Competencies and Students' Satisfaction. *International Education Studies*, *7*(1), 37–46.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innovative Higher Education*, *40*(4), 291–303.

Maricic, M., Djokovic, A., & Jeremic, V. (2016). Gender bias in student assessment of teaching performance. *Central European Conference on Information and Intelligent Systems*, 137–137.

Marsh, H. W., & Bailey, M. (1993). Multidimensional Students' Evaluations of Teaching Effectiveness: A Profile Analysis. *The Journal of Higher Education*, *64*(1), 1–18.

Meyer, J. P., Doromal, J. B., Wei, X., & Zhu, S. (2017). A Criterion-Referenced Approach to Student Ratings of Instruction. *Research in Higher Education*, *58*(5), 545–567.

Punyanunt - Carter, N., & Carter, S. L. (2017). Students' Gender Bias in Teaching Evaluations. *Higher Learning Research Communications*, *5*(3), 28–37.

Rosen, A. S. (2017). Correlations, Trends and Potential Biases among Publicly Accessible Web-Based Student Evaluations of Teaching: A Large-Scale Study of RateMyProfessors.com Data. *Assessment & Evaluation in Higher Education*, *43*(1), 31–44.

Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The Influence of Student Sex and Instructor Sex on Student Ratings of Instructors: Results from a College of Communication. *Women's Studies in Communication*, *30*(1), 64–77.

Huemer, M. (1998). Student Evaluations: A Critical Review.

Whitworth, J. E., Price, B. A., & Randall, C. H. (2002). Factors That Affect College of Business Student Opinion of Teaching and Learning. *Journal of Education for Business*, *77*(5), 282–289.