# 36-401 Data Exam 2: Instructor Quality Ratings

Anusha Bhat

2023-11-30

## 1. Introduction

For this report, we are tasked by the Vice Provost of the University of Southern North Dakota (USND) at Hoople to investigate instructor evaluations in an effort to discover the importance of evaluations in the determination of whether or not an instructor should be promoted. Although anonymous course evaluations and third-party course/instructor review websites provide relevant markers of an instructor's teaching capability, they are often susceptible to student biases. For example, students often rate instructors based on if the class is easier, if they have a gender bias in line with the instructor's gender, if they find the instructor attractive, if they like the discipline of the course, and for many more reasons. Since these evaluations are often used to improve the quality of courses a university offers, it is important to understand the influence of these biases on instructor evaluations. We hypothesize that students assign more favorable and higher ratings to instructors that positively satisfy their biases.

Ratings of 366 instructors from a college in the Midwest that were posted on RateMyProfessors.com were collected and summarized by Bleske-Rechek and Fritsch in 2011. The dataset contains average ratings and other variables for instructors with at least 10 ratings over a several year period with student ratings provided on 5 point scales. To explore the Vice Provost's overarching question, we will focus on two specific research questions. For the first question, we will investigate how quality ratings are associated with an instructor's gender, physical attractiveness, easiness of the class, and discipline. For the second research question, we will investigate whether or not the relationship between quality ratings and easiness ratings depend on the instructor's gender and discipline. Through our analysis we found that there is a positive association between quality ratings and easiness and attractiveness ratings respectively. We also found that quality ratings does depend on gender and discipline

## 2. Data

Our dataset, provided through the "alr4" package in R, contains 17 variables corresponding to various information about the instructors. Specifically, we are provided with the instructor's gender, the number of years the instructor had ratings (between 1999 and 2000), the number of ratings for an instructor, the number of courses taught by an instructor, pepper ratings of physical attractiveness, the instructor's discipline and department, average ratings for quality, helpfulness, clarity, easiness, and rater interest, and the standard deviations of these six rating types. For our purposes, we will focus on the average quality and easiness ratings, gender, pepper attractiveness, and discipline variables in our investigation.

An instructor's gender is categorically either male or female. Attractiveness is also a categorical variable with the categories yes and no. If most students rate an instructor attractive, then they receive a pepper attractiveness rating of "yes", otherwise they receive a "no". Our last categorical variable is the instructor's discipline, with the following categories: humanities, social sciences, STEM (science, technology, engineering, and mathematics), and pre-professional training. Average quality and average easiness ratings are continuous variables on a scale of 1 (worst rating) to 5 (best rating). The average quality rating corresponds to the average of the overall quality of an instructor's course, meanwhile, the average easiness rating corresponds to the average easiness of an instructor's course(s). Before we conduct our investigation, we will perform some exploratory data analysis on our variables of interest.

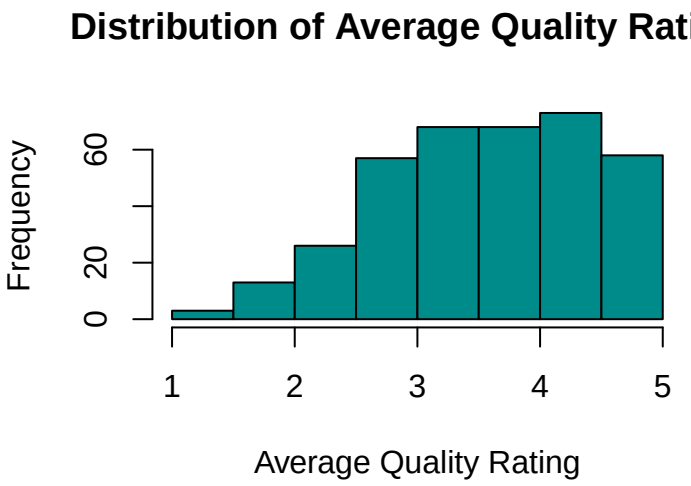


Figure 1: The average quality ratings has a left skewed distribution, with a range from 0-5. Most instructors received a rating in the 3-5. Instructors received a rating of 1 the least frequently.

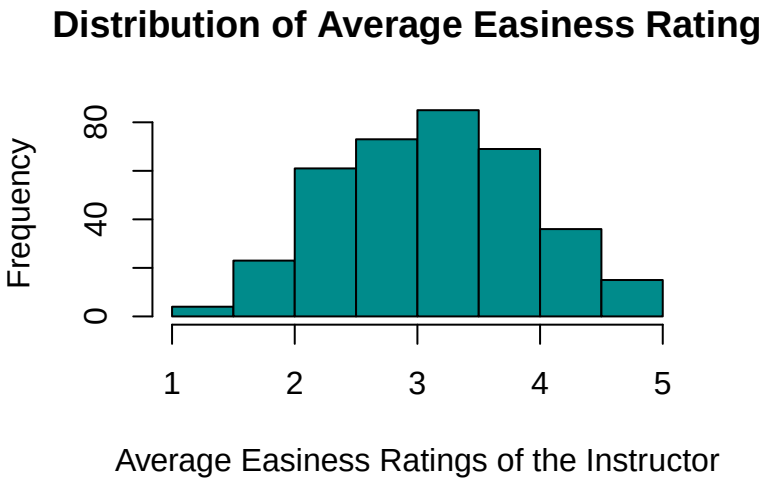


Figure 2: The average easiness ratings range from 1-5, with a median rating between 3-3.5. Most professors received a rating of 2.5-4. Instructors received a rating of 1 the least frequently.

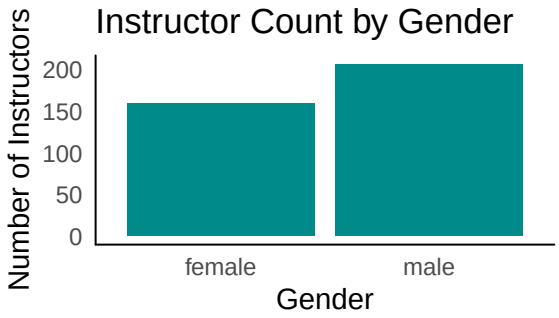


Figure 3: There are more male instructors than female instructors.

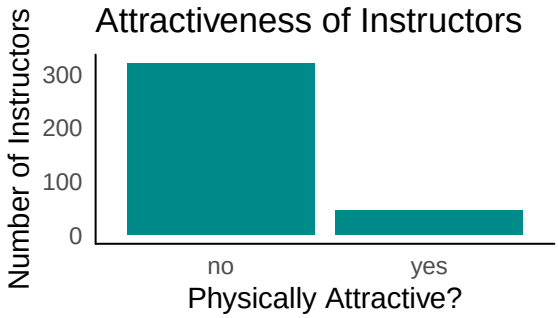


Figure 4: Most instructors received a chili pepper attractiveness rating of 'no'.

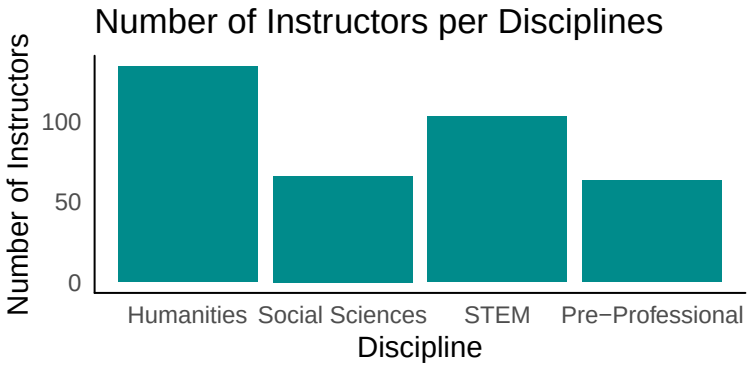


Figure 5: Humanities courses have the highest number of instructors, followed by STEM courses. Pre-professional has the lowest number of instructors, falling slightly shorter than the social science courses.

We note that we performed a log transformation of the variable, however, this did not improve the skewed shape so we will proceed without applying any transformations to the variable, but note that this may impact our linear regression results. Average easiness ratings has a bell curve shape, so, we will not apply any transformations to the variable. Additionally, the drastic difference in the number of attractive and unattractive instructors may influence our findings.

Next, we will look at bivariate elementary data analysis.

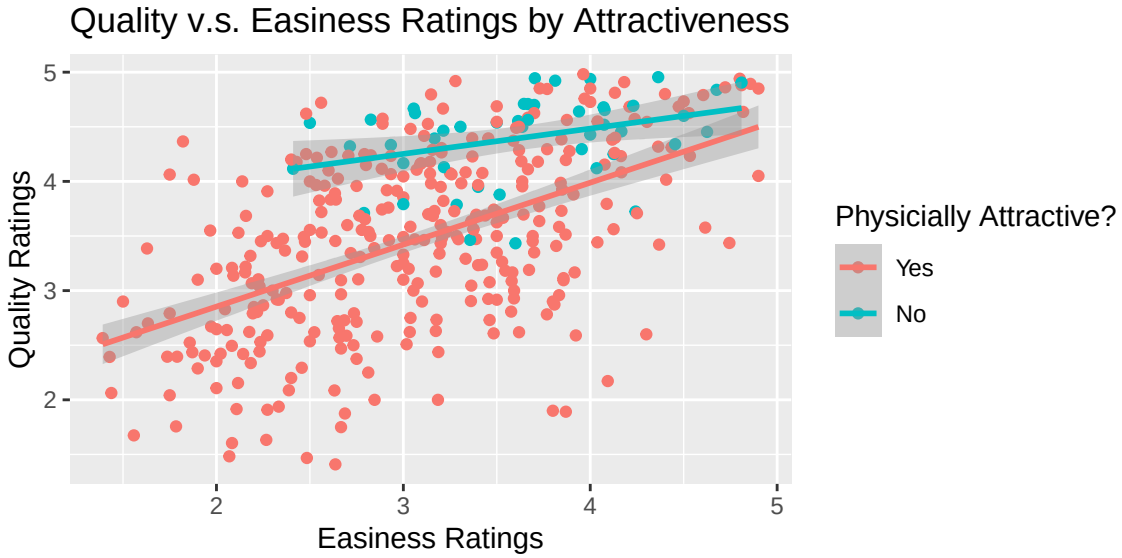


Figure 6: The confidence intervals for the regression lines for each attractiveness rating do not overlap, indicating that the slopes may be different between the lines. There is a positive linear relationship between easiness and quality ratings.

Due to the linearity observed in figure 6, we can use a linear regression model for our analysis. It also appears that unattractive professors cluster towards higher easiness and quality ratings compared to attractively rated professors. Scatter plots for the models quality ~ easiness + gender and quality ~ easiness + discipline also yield similar results of overlapping regression lines. Additionally, we found a similar proportion of attractiveness ratings for both genders and a similar proportion of attractiveness ratings for each discipline. However, we found that there was not a similar proportion of instructor genders for each discipline, highlighted in figure 7.

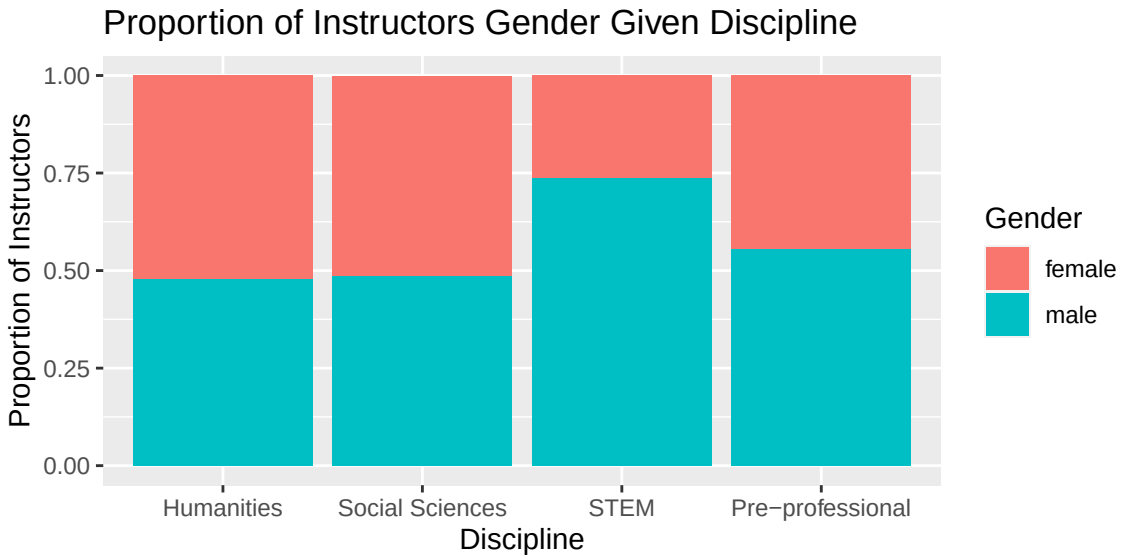


Figure 7: STEM courses have the highest proportion of male instructors compared to the proportion of male instructors for the other disciplines.

## 3. Methods

# 3.1 Research Question 1

To address the first research question regarding how the quality rating is associated with an instructor's easiness rating, pepper attractiveness rating, discipline, or gender, we can construct a linear model with these quality ratings as the response variable and the remaining four variables as predictor variables. Additionally, since we found in figure 6 that the lines have different slopes, we will include an interaction term between easiness and attractiveness in our model. Since we did not find different slopes for predicting quality ratings with the other categorical variables, we will not include interaction terms between them and

easiness. We also found in figure 7 that there is a higher proportion of male instructors in STEM courses compared to the other disciplines, so we will include an interaction term between gender and discipline. Since we found similar proportions of attractiveness ratings for gender and for discipline, we will not include an interaction term between gender and attractiveness or between gender and discipline. This model will allow us to observe any potential relationships between the predictor variables and the response. Thus, our model will be:

$$\text{Quality} = \beta_0 + \beta_1 \text{Easiness} + \beta_2 \mathbb{1}_{Attractiveness=yes} + \beta_3 \mathbb{1}_{Gender=male} + \beta_4 \mathbb{1}_{Discipline=socsci} + \beta_5 \mathbb{1}_{Discipline=STEM} + \beta_6 \mathbb{1}_{Discipline=preprof} + \beta_7 \text{Easiness} \mathbb{1}_{Attractiveness=yes} + \beta_8 \mathbb{1}_{Gender=male} \mathbb{1}_{Discpline=socsci} + \beta_9 \mathbb{1}_{Gender=male} \mathbb{1}_{Discpline=STEM} + \beta_{10} \mathbb{1}_{Gender=male} \mathbb{1}_{Discpline=preprof} + \epsilon_i$$

We can perform $t$-tests (for 355 degrees of freedom and an alpha level of 0.05) with the null hypothesis that $\widehat{\beta}_i = 0$ for each $\widehat{\beta}_i$ in our model. Our alternative hypothesis is that $\widehat{\beta}_i \neq 0$. If we reject the null hypothesis for a parameter corresponding to a particular variable or interaction, then we can conclude that there is an association/relationship between the variable or interaction and quality ratings. We can further construct 95% confidence intervals for any parameters we may find to be statistically significant, in order to describe the association between that predictor variable and quality ratings. For the $t$-test and the confidence intervals, we will assume that the errors follow a multivariate Gaussian distribution with constant variance and mean 0, allowing us to further assume that all the beta parameters in our model are normally distributed as well for valid confidence intervals. Additionally, we also assume that the data are independent and randomly sampled. We can check if these assumptions are valid using a residual v.s. fitted values plot and a QQ plot.

We note that residuals v.s. predictor plots for all the predictor values had no noticeable issues or trends in the errors. Since there are no trends in the plots from figure 8, our assumptions are reasonably satisfied. Although there are some outliers in the residuals v.s. fitted values plot and the residuals do not fall perfectly onto the normal-line for the QQ plot, we will proceed with our analysis since there are greater than 30 observations in our data. However, we will note that this may be a limitation for our confidence intervals since we assume that the errors are normally distributed in order to further assume that all $\widehat{\beta}_i$ parameters are normally distributed as well.

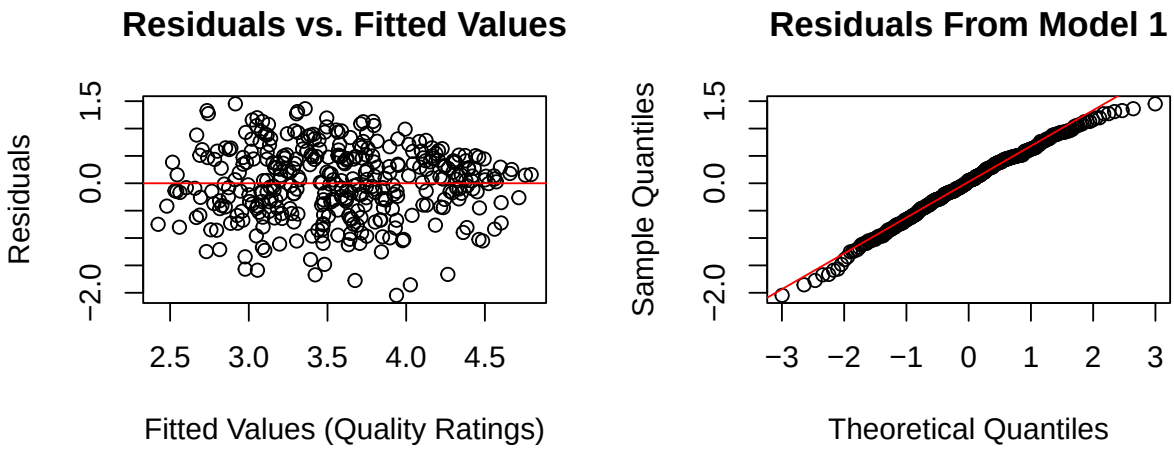


Figure 8: The residuals are randomly scattered around the line y = 0 and appear to have mostly constant variance. There are a few outliers in the range of residuals greater than 1 and less than -1.5. The residuals fall approximately along the red line for the QQ plot, indicating that they follow a multivariate Gaussian distribution.

## 3.2 Research Question 2

To answer our second research question, we can construct a nested F-test using a full and reduced linear regression model and an alpha level of 0.05. The full model will include all of our predictor variables and the reduced model will exclude a subset of these parameters. Our full model is:

$$\text{Quality} = \beta_0 + \beta_1 \text{Easiness} + \beta_2 \mathbb{1}_{Attractiveness=yes} + \beta_3 \mathbb{1}_{Gender=male} + \beta_4 \mathbb{1}_{Discipline=socsci} + \beta_5 \mathbb{1}_{Discipline=STEM} + \beta_6 \mathbb{1}_{Discipline=preprof} + \beta_7 \text{Easiness} \mathbb{1}_{Attractiveness=yes}$$

We have omitted the interaction terms between gender and discipline in our full model since we want to observe the dependency of quality ratings on gender and discipline rather than their interaction terms. The reduced model will omit the parameters for gender and discipline, i.e.:

$$\text{Quality} = \beta_0 + \beta_1 \text{Easiness} + \beta_2 \mathbb{1}_{Attractiveness=yes} + \beta_7 \text{Easiness} \mathbb{1}_{Attractiveness=yes}$$

Our null hypothesis is that $\beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ with the alternative hypothesis being that at least one of $\beta_3, \beta_4, \beta_5, \beta_6$ does not equal 0. If we reject the null hypothesis, then we can conclude that quality ratings does depend on gender and discipline since we need them in our full model when predicting quality ratings. In the event that we reject the null hypothesis, we can conduct *t*-tests on the gender and discipline coefficient estimates in the full model (for 358 degrees of freedom and an alpha level of 0.05) to determine the nature

of this dependence.

To conduct the nested f-test, we must assume that the errors follow a multivariate Gaussian distribution. Since the reduced model is a subset of the full model, these assumptions on the full model's errors will apply to the reduced model's errors as well. To conduct the $t$-test we must assume that the errors follow a multivariate Gaussian distribution with constant variance and that the data are independently and randomly sampled. We can again identify if our models satisfy these assumptions using the same diagnostic techniques as we did for the previous research question. Since the full model is a superset of the reduced model and we may use it for the $t$-tests, it is sufficient to only display the full model's diagnostic plots.

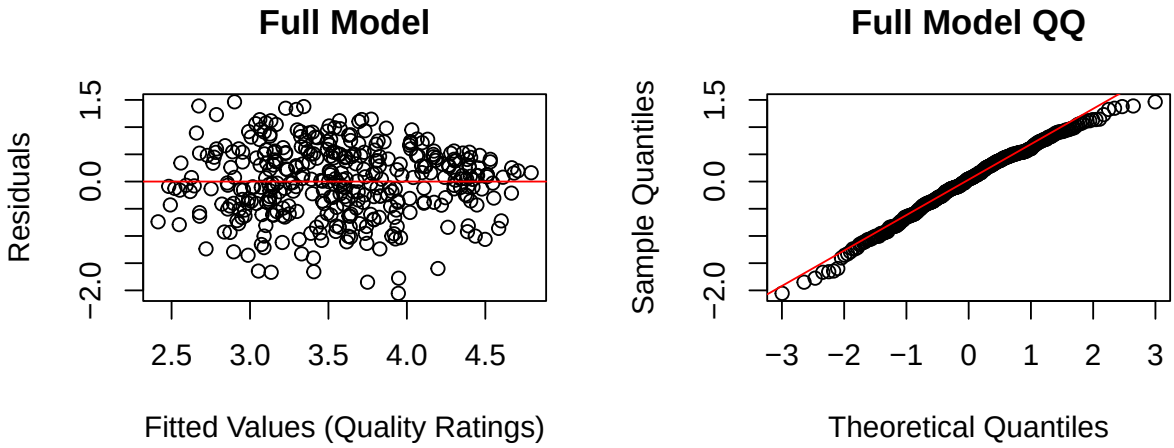


Figure 9: The residuals are randomly scattered around the line y = 0 and appear to have mostly constant variance. There are a few outliers in the range of residuals greater than 1 and less than -1.5. The residuals fall approximately along the red line on the QQ plot, indicating that they follow a multivariate Gaussian distribution.

We again note that the residuals v.s. predictors plots did not display any trends in the residuals. Since there are no trends in the plots from figure 9, our assumptions are reasonably satisfied. Despite the outliers in the residuals v.s. fit plot and the residuals not falling perfectly on the normal-line in the QQ plots, we will proceed with our analysis since we have greater than 30 observations–noting the same limitations as previously mentioned.

## 4. Results

In this section, we will analyze the models and hypothesis tests we constructed to address both of our research questions. We will first analyze model 1 to answer the first research

question, then we will conduct our nested F-test to answer our second research question.

## 4.1 Research Question 1

The estimates of the coefficients of the parameters for model1, as well as their respective standard errors, $t$ statistics and p-values are displayed in table 1. This model has an estimated variance of 0.654 and a residual squared error of 0.4069.

Table 1: Coefficients for Model 1

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 1.4889 | 0.1691 | 8.8058 | 0.0000 |
| easiness | 0.5999 | 0.0489 | 12.2706 | 0.0000 |
| attractivenessyes | 1.9907 | 0.6330 | 3.1447 | 0.0018 |
| gendermale | 0.1279 | 0.1134 | 1.1278 | 0.2602 |
| disciplineSocSci | 0.0814 | 0.1376 | 0.5919 | 0.5543 |
| disciplineSTEM | 0.1267 | 0.1489 | 0.8508 | 0.3954 |
| disciplinePre-prof | -0.0936 | 0.1482 | -0.6314 | 0.5282 |
| easiness:attractivenessyes | -0.3750 | 0.1733 | -2.1641 | 0.0311 |
| gendermale:disciplineSocSci | -0.1333 | 0.1979 | -0.6737 | 0.5010 |
| gendermale:disciplineSTEM | 0.0786 | 0.1853 | 0.4242 | 0.6717 |
| gendermale:disciplinePre-prof | 0.1617 | 0.2032 | 0.7955 | 0.4269 |

From the $t$-tests for each coefficient with the null hypothesis that the coefficient equals zero and an alternative hypothesis that it does not equal zero (for an alpha level of 0.05), we find that only $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_7$ have a statistically significant association with quality ratings. In particular, with all other variables held constant, a one unit increase in easiness ratings is positively associated with an increase of approximately 0.6 units, on average, in quality ratings for instructors who are female, not attractive, and teach humanities courses ($t(355) = 12.2706$, 95% CI[0.504, 0.696], p-value $< 0.01$). In contrast, we find that for female instructors who are attractive and teach humanities courses, a one unit increase in easiness ratings is positively associated with an increase of 0.225 units, on average, in quality ratings with all other variables held constant ($t(355) = -2.1641$, 95% CI[-0.716, -0.034], p-value = 0.0311). In addition, for instructors who are female, not attractive, and teach humanities, we find that quality ratings are approximately 1.49 when the instructor receives an easiness

rating of 0 (t(355) = 8.8058, 95% CI[1.156, 1.822], p-value < 0.01). This baseline value of quality ratings is approximately 3.4 when an instructor receives a 0 easiness rating for the course but is instead attractive (rather than not) for female instructors who teach humanities courses (t(355) = 3.1447, 95% CI[0.746, 3.236], p-value = 0.0018).

We find that the interaction coefficients between gender and discipline are not statistically significantly associated with quality ratings and neither are the coefficients for gender and discipline. We can conclude that the magnitude of change in quality ratings for a one unit increase in easiness ratings is the same regardless of instructor gender or discipline but will change depending on attractiveness. We can investigate this further to see whether or not quality ratings may truly depend on gender and discipline in the next subsection.

## 4.2 Research Question 2

Table 2 displays the results of our nested F-test. The first row corresponds to the reduced model and the second row corresponds to the full model. We note that our exclusion of the interaction terms between gender and discipline in our full model is further supported by our findings in the previous subsection that these interaction terms are not statistically significantly associated with quality ratings.

Table 2: Results of the Nested F-Test

| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|
| 362 | 157.6491 | NA | NA | NA | NA |
| 358 | 152.6132 | 4 | 5.0359 | 2.9533 | 0.0201 |

The nested F-test is conducted with the null hypothesis that the reduced model is correct and the alternative hypothesis that the full model is correct (i.e., at least one of the coefficients for gender or are statistically significantly different from zero). We reject the null hypothesis and find that an instructor's quality rating does depend on gender and discipline in this model (F(4, 358) = 5.0359, p-value = 0.0201). We can further investigate this dependence with a *t*-test on the coefficient estimates for gender and discipline in our full model (for 358 degrees of freedom and an alpha level of 0.05). For each coefficient, we will use the null hypothesis that the the estimate $\widehat{\beta}_i = 0$ versus the alternative hypothesis that $\widehat{\beta}_i \neq 0$. Table 3 displays the results of this test on the full model, with coefficients omitted for non gender and discipline related variables.

Table 3: Results of the Full Model

| | term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|---|
| 1 | (Intercept) | 1.4830 | 0.1665 | 8.9050 | 0.0000 |
| 4 | gendermale | 0.1500 | 0.0708 | 2.1189 | 0.0348 |
| 5 | disciplineSocSci | 0.0176 | 0.0986 | 0.1786 | 0.8583 |
| 6 | disciplineSTEM | 0.1800 | 0.0891 | 2.0206 | 0.0441 |
| 7 | disciplinePre-prof | -0.0049 | 0.1005 | -0.0486 | 0.9612 |

In particular, we find that a male instructor who is unattractive and teaches a humanities course has a quality rating 0.15 units higher than a female instructor who is unattractive and also teaches a humanities course when they both receive an easiness rating of 0 ($t(358) = 2.1189$, 95% CI [0.010778587 0.28916427], p-value = 0.0348). Additionally, a female instructor who is unattractive and teaches a social science course has a quality rating approximately 0.018 units higher than a female instructor who is also unattractive but teaches a humanities course instead when they both receive an easiness rating of 0 ($t(358) = 0.1786$, 95% CI[-0.1763, 0.2116]). We observe that there is a difference in quality rating of an instructor when their easiness rating is 0 depending on if they are male instructor (teaching any discipline) or if they are a female instructor who teaches social science. Thus, we can conclude that quality ratings do depend on gender and discipline. We note that this finding should be further investigated since it is possible that with a different set of full and reduced models, we may find that quality ratings may not depend on gender and discipline.

## 5. Conclusions

To summarize our investigation, we found that the easiness ratings and attractiveness are positively associated with quality ratings respectively. In particular, female instructors that teach humanities with higher average easiness ratings may have higher average quality ratings as well. We note that for this observation, the magnitude of the increase in quality ratings for a one unit increase in easiness ratings is greater if the instructor is unattractive compared to attractive. We also found that a female instructor who is rated attractive, taught a humanities course and has an easiness rating of 0, tends to have a quality rating higher than a female instructor who was rated as unattractive but also taught a humanities course and received an easiness rating of 0. Lastly, we found that quality ratings does depend on gender and discipline of the instructor since different genders and disciplines have different quality

ratings when they have an easiness rating of 0. These findings match the theory that student reviews of instructors are often biased.

We recommend that the Vice Provost should consider course evaluations with a grain of salt. How students rate a course is somewhat indicative of an instructor's quality of teaching so it should be taken into consideration, however, due to personal biases the student ratings may not be completely accurate. For example, a student may rate an instructor very poorly but another student may rate that same instructor highly. Thus, when deciding whether to promote an instructor, the Vice Provost should take into consideration the student ratings but should also consider other metrics (e.g. grade averages in the courses taught or average numbers of hours that student report dedicating to the course) before coming to a concrete decision.

Despite this recommendation, we should note the limitations on the generalizability of our analysis. Firstly, there was a significantly higher amount of instructors rated unattractive than the amount that were rated attractive which could influence our results since this is not reflective of a population of instructors that may have an even distribution of attractiveness. Furthermore, there could be confounding variables influencing instructor quality ratings that were not included in our investigation such as the number of years an instructor has been teaching. Often times, an instructor with more years of teaching experience has a better overall quality of instruction. Additionally, the dataset was taken from a college in the Midwest so our findings can only be applied to the instructors at that particular college. A similar study should be replicated with instructors from the University of Southern North Dakota at Hoople to provide the Vice Provost with more accurate recommendations.

There are several ways that we could improve our analysis. In the future, we could begin by collecting similar data on a larger number of instructors from a wide variety of universities in the USA. This will allow us to get more representative findings that highlight gender trends in higher education instruction. We could also collect data on more quantitative variables (e.g. time duration of one class period, length of course) and categorical variables (e.g. faculty appointments/positions, highest degree of the instructor, whether they teach undergraduate or graduate students). This will provide us with more information that can lend itself to a stronger analysis.

Future research can be done to examine how instructor quality ratings vary among instructors with different degree levels, those who teach undergraduate versus graduate level classes, as well as instructors who have different graduate degrees. These factors also influence the difficulty of a class which in turn influences quality ratings, warranting further investigation.