The Inherent Biases Effecting the Evaluation of College Courses

Sydney Benson, Josh Thyen, Jeff Courneya and Jill Wanner

STAT 333

1 Introduction

Each semester, college students around the United States provide feedback and reviews for their professors and courses. These ratings can have potential ramifications such as salary raises, future teaching opportunities, course restructuring, and tenure status (Neumann, 2000). Due to the numerous areas that these evaluations impact, it is critical to receive honest and fair ratings from students. One of the implicit assumptions made when using student-based ratings is that these evaluations reliably and accurately score professors based strictly on their teaching ability. However, humans are biased creatures, with entire branches of statistics and psychology dedicated to studying inclinations and preferences. In examining research conducted on this topic, it appears that it is human nature to be predisposed to like people more when they are young and attractive (Buck & Tiene, 1989). Based on this, we hypothesize that these characteristics will influence the evaluation scores given in courses. Prior studies of professor evaluations have indicated that factors such as age (Wilson, Beyer, & Monteiro, 2014), gender (Smith, Yoo, Farr, Salmon, & Miller, 2007), and the physical attractiveness of the professor (Goebel & Cashen, 1979) do impact evaluation scores.

In 1979, Goebel and Cashen conducted a study where images of teachers were shown to students in grades two, five, eight, eleven, and thirteen. Based on these pictures, the students were asked to sort the teachers into age and attractiveness categories. They were later presented the pictures in a random order and asked to rate the teachers based on a seven-question survey. After analyzing the results across all developmental ages, it was determined that the factors of age, sex, and attractiveness were significant factors that affected student ratings of teachers. Their statistical analysis indicated a trend towards lower ratings for middle-aged unattractive females and older unattractive males. (Goebel & Cashen, 1979). A later study done at the University of Texas reached a similar conclusion, stating that their findings indicated that "measures of perceived beauty have a substantial independent positive impact on instructional ratings by undergraduate students" (Hamermesh & Parker, 2005).

Smith et al. (2007) examined the influence that the sex of both the professor and student had on evaluations. For this study, the researchers used the results from the standard evaluation form used at the university, which had the students evaluate instructors based on five different criteria: instructor involvement, student interest, student-instructor interaction, course demands, and course organization. The evaluation used by this university asked students to score professors on a scale of 1 to 5 in a variety of different categories, with 1 indicating an outstanding score and 5 being a poor score. The results of this study concluded that students rated female professors higher in areas relating to the classroom learning environment, with the mean evaluation score for a female faculty member being 2.03 compared to 2.22 for males (p-value < 0.01) (Smith et al., 2007).

A more recent study conducted at the University of Southern Georgia by Wilson, Beyer, and Monteiro focused on the effect of age on professor evaluations and found that age was a factor in teaching evaluations for both male and female professors. Using multivariate testing, they found a significant difference in evaluation scores when examining the effect of gender, age, and the interaction between gender and age. Based on their findings, this study concluded that student evaluations are affected by inherent attributes of the teacher such as age, gender, and attractiveness (Wilson et al., 2014).

As evidenced by these studies, there exist several factors outside of teaching ability that impact professor evaluation ratings. Due to how extensively these evaluation scores are used throughout the academic process, we wish to expand off these prior findings and further establish the relationship between course ratings and the preceding factors. We also examine whether the proportion of students who submitted an evaluation for their course affects the course's assessment. This was based on personal experience, as we hypothesized that students more frequently participate in course surveys when they had either a very positive or negative experience.

Therefore, the purpose of this study was to corroborate previous studies' findings, as well as quantify the factor by which a professor's age, gender, physical appearance, and the proportion of students who submitted an evaluation for their course effect the course's overall evaluation score using multiple linear regression.

2 Methods

2.1 The Data Set

The data set we worked with came from the study completed by Hamermesh and Parker (2005). The goal of their study was to determine the relationship between physical attractiveness and teaching productivity. The data was collected from the University of Texas at Austin. The variable we aimed to predict was the overall evaluation score for a course. The observations for this variable were collected from the evaluations voluntarily completed by students in each

of the courses in our sample sometime within the last three weeks of a 15-week semester. The evaluations were administered by a student while the instructor was absent from the classroom. Each evaluation score for a course is an average of the responses given by students to the statement "Overall, this course was very unsatisfactory (1); unsatisfactory (2); satisfactory (3); very good (4); excellent (5)." Each qualitative rating of the class is shown next to its corresponding quantitative rating.

The sample of 463 courses chosen for this study was taken from all departments within the university and the course evaluation scores given are taken from courses taught between 2000 and 2002. Each instructor is represented in the sample between 1 and 13 times, depending on how many of their courses are in the sample so, although the courses themselves are independent, the instructors of the courses are not independent. For this reason, we reinforce the notion that the evaluation scores used for our study are an evaluation of the course.

The second variable we used in our study is the proportion of students in each course that submitted an evaluation of the course. In the original data set, the information for this variable came from two separate data columns. The first being the number of students who submitted a course evaluation, ranging from 5 to 380 students per course, and the second being the number of students in each course, ranging from 8 to 581 students. Thus, the proportion of students in each course that submitted an evaluation of the course ranged between 0.104 and 1.

The third variable we looked at in this study is a rating of each instructor's physical appearance. Photos of each instructor were obtained from their department's website and the photos were then rated on a scale of 1 (least attractive) to 10 (most attractive) by a panel of six students. This panel consisted of three men and three women, with two of each gender being juniors or seniors and one of each gender being a freshman or sophomore. The beauty ratings given by each student were then normalized, and the six normalized ratings were summed for each instructor so that the range of instructor beauty ratings was –1.45 to 1.97.

The final variables that we chose to focus on in this study were the age and gender of the instructor.

2.2 The Analysis

The first model we built used the instructor's age and gender, the proportion of students in a course who submitted a course evaluation and the instructors' normalized physical attractiveness score to predict the overall course evaluation score. We refer to this as the full model. Since we used multiple predictor variables in our model, we first checked for any multicollinearity in our variables. In the event that we had detected high levels of correlation (greater than 0.9) between any two variables or found a variable that had a variance inflation factor greater than 10, we would have revised the variables chosen for our study.

Then, we used a residual plot to check the model for any evidence of heteroscedasticity in the data. In the event of any heteroscedasticity, we tried several different transformations on the data. The attempted transformations included logarithmic, square root and inverse transformations given that the variable was deemed appropriate for the technique. We could also detect any possible outliers in our data using the residual plot.

Next, we looked at a QQ-plot for our model to ensure that our residuals followed a normal distribution. Then, we interpreted our scale-location plot. This gave us more evidence of whether a trend in our residuals existed. Finally, we looked at the leverage plot. This plot identified any points isolated from the rest that have too great of an influence on our model.

Once we were assured that our model had an appropriate mean function, we looked at the significance of the predictor variables in the model, using a T-test with a significance level of 0.05, to determine whether any of them were unnecessary. We also looked at the coefficient of determination to understand how much of the variation in overall course evaluation score was accounted for by our predictor variables.

In the event that one or more of our predictor variables proved to be insignificant in predicting overall course evaluation score, we removed that variable from our model. We repeated the process of checking the model assumptions and making any necessary transformations so that we could interpret the results of our model and be assured that they accurately reflected what the data had to tell us.

The final model we were interested in was the model including an interaction term between age and gender. We called this our interaction model. We kept all other variables deemed to be significant in our previous model in this model. Similar to the other models, we needed to produce various diagnostic plots to check the model assumptions, including a residual plot, QQ-plot, scale-location plot and leverage plot. Additionally, we tried transformations on this model providing that there was evidence of heteroscedasticity. After the mean function was finalized, we determined the significance of the predictors using a T-test with our same significance level. We also looked at both confidence intervals and point estimates for our coefficients to determine the effect of each predictor variable.

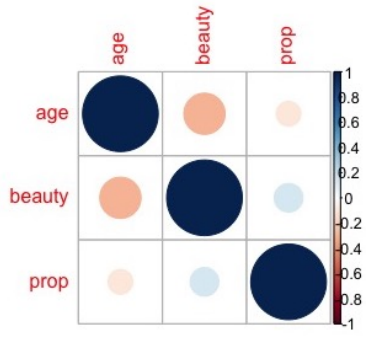


Figure 1: Correlation plot of quantitative variables included in the full model.

In order to determine the optimal model to predict overall course evaluation score, we used the Akaike information criterion (AIC) to compare the performance of each model. To choose a more complicated model, we required that the AIC be improved by at least 10 units over the simpler model.

3 Results

3.1 Multicollinearity

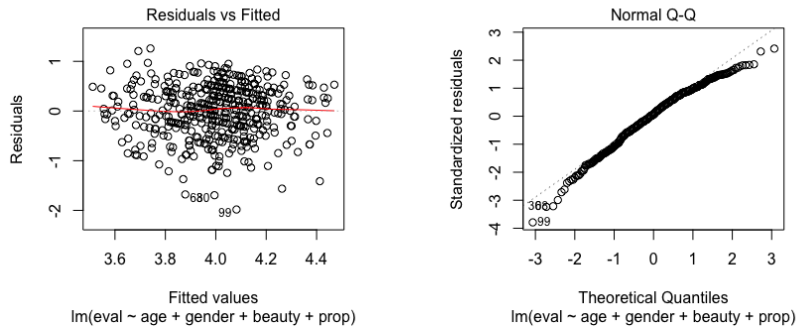
To understand the correlation of our variables we used two different methods. First, we looked at the correlation between pairs of the quantitative predictor variables. From Figure 1 we can see that all but one of the correlation coefficients was negative. The two variables that had a positive correlation coefficient were the proportion of students who took the survey and the beauty rating, but all of the correlation coefficients were low enough that there was no cause for concern. The second method for understanding the correlation between our variables was calculating the variance inflation factor (VIF) for all four predictor variables. All of these VIF values were near one, with the largest value being the one corresponding to age, but still being relatively small at 1.181. With all VIF factors below 1.5 and lacking any strong correlation coefficients, we concluded that there was no obvious multicollinearity present within the predictor variables.

3.2 The Full Model

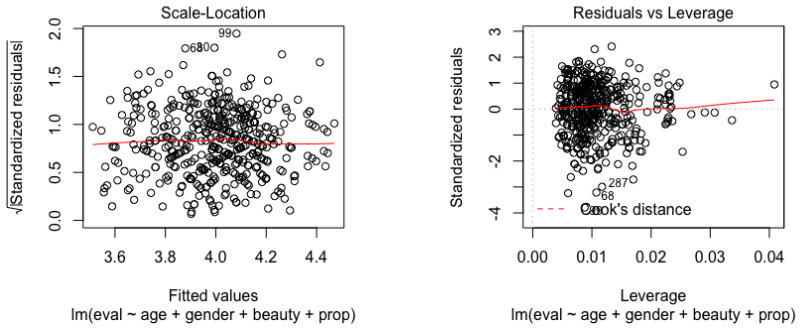
Again, we refer to our full model as the model using the instructor's age and gender, the proportion of students in a course who submitted a course evaluation, and the instructors' normalized beauty score to predict the overall course evaluation score. The residual and scale-location plots (Figures 2a and 2c) illustrate that our model had a minor issue with over-predicting a small number of course evaluation scores. The over-prediction occured for both low and high evaluation scores. There also appeared to be larger variance around an evaluation score of 4.0. This may have occured because professors who receive a mid-level course evaluation score vary more both in terms of personal characteristics and how students view them and the course than professors who receive both low and high course evaluation scores. We attempted to improve this slight heteroscedasticity by performing various combinations of logarithmic, square root, and inverse transformations on the appropriate variables. There was not any notable improvement to the slight heteroscedasticity violation. Therefore, we deemed the added complexity in explaining the results of a model utilizing these transformations not useful in the scope of this discussion.

The QQ-plot (Figure 2b) indicated that this model does not violate the normality assumption of the residuals severely. The standardized residuals mostly followed the straight line indicating that they come from a normal distribution with slight curvature for the lower and upper quantiles. The leverage plot (Figure 2d) indicated that there were no outliers outside Cook's distance and thus there were not any overly influential outliers in this model.

After determining that the model assumptions were satisfied and deciding not to apply any transformations, we built the multiple linear regression model.



(a) Residual plot (b) QQ-plot



(c) Scale-location plot (d) Leverage plot

Figure 2: Diagnostic plots for the full model.

The regression equation is:

$$\widehat{evaluation} = 3.663 - 0.002age - 0.225female + 0.122beauty + 0.703proportion.$$

The regression coefficients for beauty and proportion are both positive, while a higher age is predicted to lower the course evaluation score. The gender coefficient indicated that evaluation score decreases by 0.225 when the instructor is female and all other variables are held constant. Figure 3 illustrates the point estimates and confidence intervals for the regression coefficients at a confidence level of 95%. Thus, we are 95% confident that the true values of these coefficients lie within their respective intervals when all other coefficients are held constant.Exact values for the confidence intervals can be found in Table A.1, in Appendix A.

The T-tests for the age, gender, beauty, and proportion predictors result in age being the only predictor in this model for which there is not enough evidence to say that the predictor is significant in the regression relationship. Table A.1

lists the p-values from each of these tests. Since age is not a significant predictor, we next attempted to build a better model by excluding it and then examining how the results of our T-tests change.

3.3 The Reduced Model

Since our full model, which included instructor age, gender, beauty rating and the proportion of students who submitted an evaluation, indicated that instructor age was an unnecessary predictor for the model, we decided to remove instructor age as a predictor from the model and determine whether this new model improves upon the model containing all predictors. Thus, this model contained instructor gender and beauty rating and the proportion of students who submitted a course evaluation as predictor variables. We refer to this second model as the reduced model.

After examining the residual plot, we found that this model has similar heteroscedasticity problems to our full model. Therefore, we tried various transformations on the reduced model. Unfortunately, none of the transformations appeared to improve this issue significantly. Due to this, we decided that the transformations which did slightly improve the model did not make large enough improvements to warrant the more difficult interpretation of predictions to include the transformations in our model.

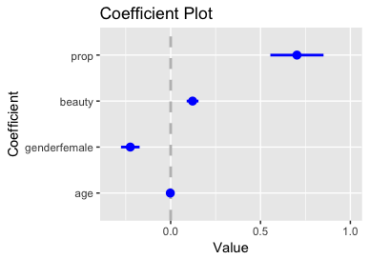


Figure 3: Coefficient plot with confidence intervals for the full model.

Regarding the other diagnostic plots, the QQ-plot, scale-location plot, and leverage plot, each of these plots appeared to be similar to the diagnostic plots of our full model. Thus, we concluded that our assumptions were met at least as well for this model as they were for our full model.

Next, we turned our attention to the regression coefficients. In this model, without instructor age as a predictor, all of our predictors were deemed significant (Table A.2). The largest p-value obtained for any predictor in this model was for instructor beauty rating and had a value of $5.56 \cdot 10^{-5}$. All of our predictors would be found significant under any small significance level and especially under our chosen significance level of 0.05. The gender predictor still has a negative coefficient, meaning female professors receive lower evaluation scores than their male counterparts on average, and the two other predictors see their coefficient values increase slightly. Looking at Figure 4, we

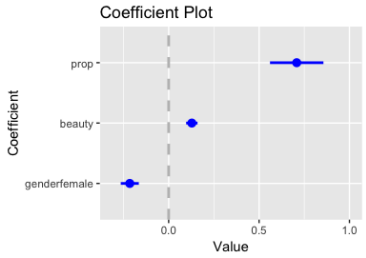


Figure 4: Coefficient plot with confidence intervals for the reduced model.

see that the widest confidence interval is for our variable corresponding to the proportion of students in a course to submit a course evaluation. The coefficient plot also illustrates that none of our 95% confidence intervals contain 0, another indicator that all of our predictor variables are significant. More detailed information for these confidence intervals can be found in Table A.2.

3.4 The Interaction Model

The next step was to analyze the model and include an interaction term with the goal of obtaining a more accurate model that would eliminate any heteroscedasticity present. The inclusion of an interaction term within a regression model can be beneficial to the mean function of a model. If two or more predictors appear to have an obvious relationship, it would be important to have the interaction term to account for that relationship. Effective interaction terms can also rid a model of heteroscedasticity. If it can provide the model a consistent variance throughout its residuals, then that would be a sufficient reason to include the interaction term.

Looking at past research, we hypothesized that age and gender would have the most significant interaction term of all possible combinations. We believed that younger female professors might show consistently higher evaluation scores, deeming an interaction term important. The inclusion of this interaction term shifted the female coefficient to a positive value while the coefficient for the age and gender interaction term was negative. Table A.3 gives detailed information for the T-tests and values of the coefficients as well as their 95% confidence intervals. The interaction term between gender and age had a p-value of 0.011, deeming it significant, but the additional term did not improve upon the diagnostic plots for the previous models. Comparing the residual plots of the models with and without the interaction term, there did not appear to be any improvement in the minimal amount of heteroscedasticity that was present in our full model.

In the case of this model, we did see that there was a relationship between age and evaluation score, dependent on gender, but the interaction term was not successful in improving any areas of heteroscedasticity that were present in the full and reduced models. We will later look at a comparison of the models' AIC values to determine if this interaction term provided significant benefits with regards to the predictive capabilities of the interaction model.

3.5 Comparison of Models

We started out by comparing the models' coefficients of determination in order to understand how much variation in course evaluation scores was accounted for by the physical characteristics of the professors and the course rather than the teaching ability of the professor and the information taught in the course. Table 1 gives the R^2 values for each of the models mentioned previously. We note that the values appear relatively close, with the range being 0.0135 between the model accounting for the most variation and the least. In doing this

Model	Coefficient of Determination
Full	0.1118
Reduced	0.1108
Interaction	0.1243

Table 1: Comparison of coefficients of determination (R^2).

comparison, we hoped to see fairly low R^2 values in each of the models because lower values would indicate that the inherent characteristics of the professor and course have little impact on the overall course evaluation score.

Finally, to understand which model represented the data most accurately, we compared the three created models using the AIC for each model. The AIC values for each model were very similar: 724.584, 723.099, and 720.083 for our full model, reduced model, and interaction model, respectively. Since none of the AIC values for our models differed by more than 10 units from any other model, we make our model recommendation based on the number of predictor variables used in each model. Since our reduced model used the fewest number of predictors and had a comparable AIC to the other models created, we recommend that any application of these models to correct for biases in course evaluation scores utilize this model, of the three models we created.

4 Discussion

Since course evaluations are so widely used throughout the academic process, it is critical to understand and identify these extraneous factors outside of teaching ability that effect ratings. As discussed in the study conducted by Smith et al. (2007) and Neumann (2000), teaching evaluations have been one of the top sources of information on teaching effectiveness. They are used by professors, administrators, and government agencies to determine salary, tenure, teaching opportunities, and even university accreditation. Therefore, in an ideal world, the benefits that come along with exceptional teaching evaluations would be bestowed strictly on the highest class of educators. However, our findings expose alternative elements that may elevate certain professors over ones that could be more deserving.

We believed that age, gender, and attractiveness were inherent characteristics of professors that were significant factors in determining course evaluation scores, based on prior research. In addition, we included the proportion of students who submitted the evaluation, as we hypothesized that a higher proportion of students would submit the evaluation if they had either an extremely positive or negative class experience. We determined that all of the preceding factors were significant with regards to course evaluation scores, a way of evaluating a professor's teaching ability, except for age, which showed no significance. Based on our own student experience, the finding that age is not a significant predictor contradicts what we expected for predicting evaluation score. Though

several of these factors had been identified as significant by prior studies, our research plays a role in advancing the existing collection of analysis on this subject. In addition to corroborating previous findings, we included the proportion of students who submitted the course evaluation, which had not previously been used as a factor. Additionally, we used multiple linear regression, allowing us to develop an equation that could potentially model these factors in the future and be used to correct for biases in teaching evaluations.

Our results reveal that there are several areas of bias when it comes to course evaluations. The first area of bias being gender, with our research revealing that course evaluation scores decrease by 0.225 points when the professor is female. Second, with regards to beauty, professors who are perceived to be unattractive have course evaluation scores that are lower than their colleagues who are perceived as attractive. A final inherent factor that influences course evaluation scores is the proportion of students in the class who submit an evaluation. Our research indicates that a higher proportion of students submitting the evaluation is correlated with higher overall evaluation scores. Therefore, if a professor has a class where a low proportion of students submit the course evaluation, their scores are adversely effected. The reduced model which used these three factors produced an R^2 value of .1108, meaning that just over 11 percent of the variability in course evaluation scores comes from inherent characteristics and factors that the professor has no control over. A study like ours, which attempts to identify biases, would ideally produce R^2 values at or near 0. This would indicate that course evaluations are purely indicative of class experience and accurately reflect the professor's teaching abilities.

Had we had the time, it would have been optimal to collect data from the University of St. Thomas to conduct an entirely independent study. However, due to time constraints it was not feasible to do this for a semester long student project. We would have liked to include an additional variable assessing the average grade received by the students evaluating the course. Our belief is that a higher average class grade would correlate with higher class evaluations as well as the opposite holding true for a lower average grade. It would be interesting to determine whether professors who give out so called "easy A's" receive higher ratings, or if students achieve higher grades because of exceptional professors. Additionally, we are intrigued by several other factors such as tenure status, salary, and native language spoken by the professor and feel as if these variables could be possible places for future research to examine.

References

[1] Hamermesh, D. S., and Parker, A. (2005). Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity.*Economics of Education Review, 24*(4), 369-376. doi:10.1016/j.econedurev.2004.07.013

[2] Wilson, J. H., Beyer, D., and Monteiro, H. (2014). Professor Age Affects Student Ratings: Halo Effect for Younger Teachers. *College Teaching, 62*(1), 20-24. doi:10.1080/87567555.2013.825574

[3] Buck, S., and Tiene, D. (1989). The Impact of Physical Attractiveness, Gender, and Teaching Philosophy on Teacher Evaluations. *Journal Of Educational Research, 82*(3).

[4] Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., and 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.

[5] Goebel, Barbara L., and Cashen, V. M. (1979). Age, Sex and Attractiveness as Factors in Student Ratings of Teachers: A Developmental Study. *Journal of Educational Psychology, 71*(5), 646-53.

[6] Neumann, R. (2000). Communicating Student Evaluation of Teaching Results: Rating Interpretation Guides (RIGs). Assessment & Evaluation in Higher Education, 25(2), 121-134.

Appendix A

Regression Coefficients Tables

A.1 Full Model

Full Model				
Effect	Coefficient	95% Confidence Interval		P-value
Age	-0.002	-0.007	0.003	0.4759
GenderFemale	-0.225	-0.326	-0.123	$1.72 \cdot 10^{-5}$
Beauty	0.122	0.058	0.186	0.0002
Proportion	0.703	0.412	0.994	$2.73 \cdot 10^{-6}$

A.2 Reduced Model

Reduced Model				
Effect	Coefficient	95% Confidence Interval		P-value
GenderFemale	-0.215	-0.313	-0.117	$2.01 \cdot 10^{-5}$
Beauty	0.128	0.066	0.190	$5.56 \cdot 10^{-5}$
Proportion	0.708	0.418	0.998	$2.19 \cdot 10^{-6}$

A.3 Interaction Model

Interaction Model				
Effect	Coefficient	95% Confidence Interval		P-value
Age	0.003	-0.004	0.0095	0.3696
GenderFemale	0.429	-0.085	0.942	0.1014
Beauty	0.124	0.060	0.188	0.0002
Proportion	0.731	0.441	1.021	$1.02 \cdot 10^{-6}$
Age:GenderFemale	-0.014	-0.025	-0.003	0.0111