How Beauty Trumps Performance: An Examination of Instructor Evaluations at the University of Texas*

Marcin Jaczynski

20 April 2023

At the end of each semester of school, students are able to rate their professors and teachers through course evaluations. However, there are concerns about the validity and reliability of these evaluations as a measure of teaching effectiveness. This paper examines the impact of instructor beauty on course evaluations at the University of Texas between 2000 and 2002, while also considering other factors that may influence ratings. The findings suggest that beauty does have a positive effect on evaluations, as such the use of the instructor evaluations for the use of hiring, firing, and promoting instructors remains controversial.

1 Introduction

Instructor ratings have a direct impact on the validity of an instructor or professor, and may be used for hiring, firing, promotion and consideration for tenure purposes. However, if beauty directly influences the instructor's evaluation, then such a metric is not a fair representation of the instructor, and therefore should be omitted in considerations for promotions, firing and tenure. While there are a lot of factors that play into how students will perceive an instructor and what kind of rating they will give them at the end of the semester, one that is not taken into consideration many times is how beautiful they are. As such, I have chosen to try and answer "How does an instructors beauty, according to their students, affect the instructors course evaluation?" This paper seeks to examine what the trends in instructor evaluations are from the students at the University of Texas during the years 2000-2002. Among beauty, things such as age, gender, department, tenure, and how many students they were teaching are going to be taken into account and evaluated. Based on the observations from the 463 instructors, there is a correlation between the beauty of the instructor and the evaluation they

*Code and data are available at: https://github.com/jackchinski/instructor-ratings

have received. These results have important implications for how we understand the factors that shape student perceptions of instructors and how we evaluate instructor performance. The correlation between beauty and evaluation is not strong, and does not persist when separating male and female instructors, where males average a higher evaluation yet score lower in terms of beauty than females. However, other things such as being native, and the class size also have an impact on the overall evaluation of the instructor. This paper will elaborate on the data, how it was sourced and analyzed, then will explore the results, and discuss them as well as their limitations.

2 Data

2.1 Data Source and Methodology

The data collected for this paper, was sourced form an Applied Econmetrics with R (AER) library. The AER is a collection of data sets in applied econometrics research, and is widely used in the academic fields of empirical analysis of microeconomic data. The package is free to download from the Comprehensive R Archive Network (CRAN), or can be installed using the RConsole or RStudio.

2.2 Data Collection

The data set used in this paper contains data pertaining to course, instructor and student evaluation data of 463 instructors. The data was collected from the 2000-2002 academic years at the University of Texas at Austin. The data was provided by Prof. Hamermesh, and includes the students ratings of their instructors as well as their beauty rating (average from six independent judges), as well as other charasteristics pertaining to the instructor. The details for the data collection methodology of the TeachingRatings data set are unclear. However, assumptions can be made that the data was collected through anonymous teacher ratings websites such as rateMyProfessor.com or course evaluations that were done at the school. However, the exact methodology is not explicitly mentioned in the documentation of the data set or the academic literature which uses the following data set.

The data set has been previously cleaned, and so it does not require any additional cleaning, however, minor adjustments have been made, such as omitting the columns "…1" and "prof" as both of them only contained the entry number, which was already present in data frame. Additionally, as a safety check, a function was ran in order to remove any N/A values, however, the result showed that no entries were removed as a result of said function.

Each entry in the following data set contains the following details:

- Minority - factor. Does the instructor belong to a minority (non-Caucasian)?

- Age - the professor's age.
- Gender - factor indicating instructor's gender.
- Credits - factor. Is the course a single-credit elective (e.g., yoga, aerobics, dance)?
- Beauty - rating of the instructor's physical appearance by a panel of six students, averaged across the six * panelists, shifted to have a mean of zero. On a scale of -2.0 to 2.0.
- Eval - course overall teaching evaluation score, on a scale of 1 (very unsatisfactory) to 5 (excellent).
- Division - factor. Is the course an upper or lower division course? (Lower division courses are mainly large freshman and sophomore courses)?
- Native- factor. Is the instructor a native English speaker?
- Tenure - factor. Is the instructor on tenure track?
- Students - number of students that participated in the evaluation.
- Allstudents - number of students enrolled in the course.
- Prof - factor indicating instructor identifier.

2.3 Data Analysis

The data set from AER was gathered by installing the AER package (Kleiber and Zeileis 2020) in the RStudio (Team 2021) console, and then the package contents were extracted to only include the TeachingRatings data set in the "download_data" file under "/scripts". The data was then analyzed using the R programming language (R Core Team 2022). Additionally, the programming packages dplyr (Wickham et al. 2023) as well as tidyverse (Wickham et al. 2019), and (Wickham and RStudio 2021) were used to analyze the data, test and create graphs.

2.4 Advantages and Disadvantages

The following data set has both some advantages and disadvantages. In terms of advantages, the data set is really easy to obtain and use. It is available online for free, and requires 3 steps in order to gather the main TeachingRatings data set. The data set that is then obtained has also already been cleaned, as such it does not require any additional work other than picking which columns will be used. Additionally, the documentation which is provided is very thoroughly written, with each column explained and examples of how to install the package and obtain the data. However, in terms of the disadvantages, firstly the documentation that is provided, does not provide any background context on the data gathering methodology, and if there was any incentive to complete the survey. The data set also contains 463 entries, which over a two year span, does not seem like a large number of entries. Additionally, the study was performed in the year 2002, which would not qualify it as a recent study, and would likely

need to be redone to be applicable to more recent years, especially after a global pandemic in the years 2020-2022.

3 Results

3.1 Male vs. Female

The first area I explored, while trying to find what affected an instructors evaluation is the gender of the instructor. According to my initial hypothesis, if one gender scored a higher in beauty metrics , then they would also score higher in terms of evaluation. As such, I decided to create two seperate graphs in order to show the correlation between beauty and evaluation for both genders. Figure 1 shows the beauty and evaluation scores for all male instructors, and, the results show that for male instructs for every additional point of beauty rating, we would expect the evaluation to increase by c

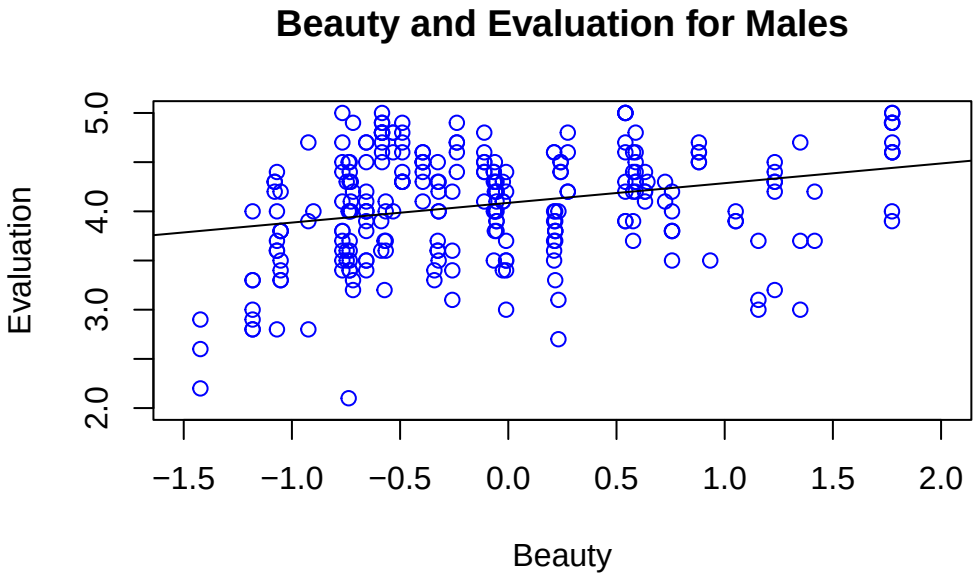

Figure 1: Beaty and Evaluation scores that increase for male instructors by 0.20027.

The next figure Figure 2 explores the correlation between the beauty and evaluation of female instructors. The figure conforms to my hypothesis, as for every additional point in beauty ratings the instructor should gain an extra 0.08762 evaluation points.

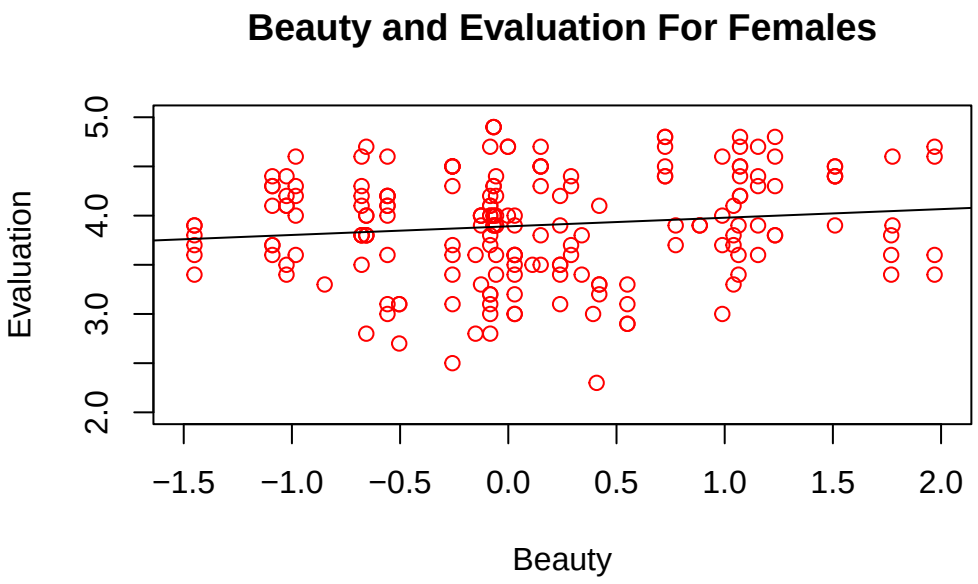

Figure 2: Beaty and Evaluation scores that increase for female instructors by 0.08762.

Figure Figure 3 shows the graph of the beauty and evaluation correlation for female instructors with the line of best fit for the male instructors. According to the models, for men the beauty affects their evaluations more than the female instructors.

Table (**table-avg-male-and-female-beauty-and-eval?**) shows the average beauty and evaluation rating for both males and females. After seeing the results of the male instructors having their evaluation be affected more by their beauty, I wanted to see what the average rating for both males and females were. According to the data, male instructors have a lower average beauty rating than female instructors, yet males are on average more highly rated. After seeing the following results, I decided to take on another route to figure out what impacted the instructors evaluation the most.

Table 1: DON'T FORGET

Gender	Beauty	Eval
Male	-0.08448224	4.069030
Female	0.11610907	3.901026

According to (**figure-overall-beauty-and-eval?**) there is a correlation between the beauty

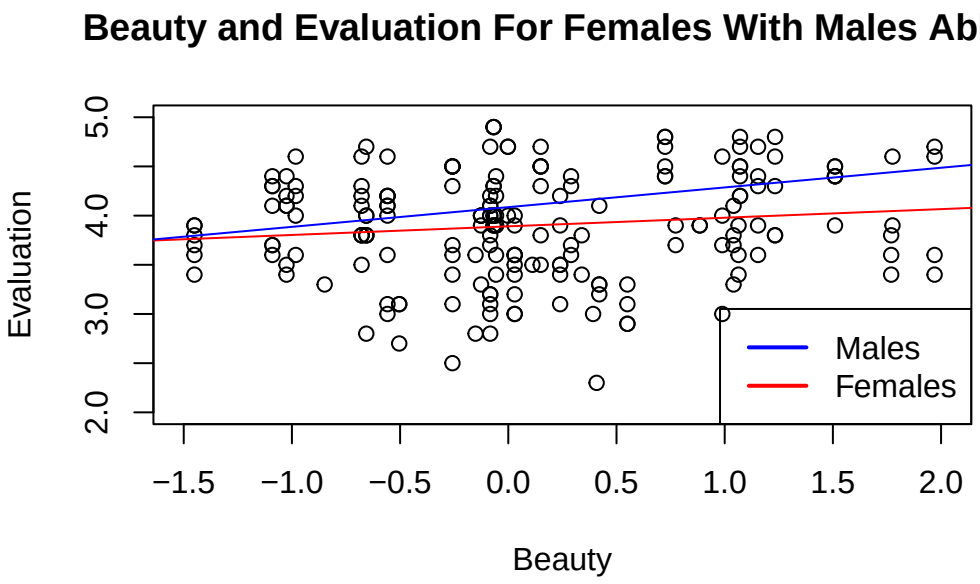

Figure 3: Female instructors beauty and evaluation with the male and female line of best fit

of the instructor and their evaluation. However, what had the biggest impact? The slope of the line for both male and female instructors is 0.133, meaning that every increase in beauty should yield an evaluation higher by 0.133. Additionally the r^2 value is 0.03364 or 3% , indicating that evaluation can be explained by the variability in beauty, which is the variability of y that is explained by the x. The r^2 indicates how much it affects the evaluation.

The table (**coefficients-affecting-eval?**) shows off all of the coefficients and how they respectively affect the evaluation of the instructor. The p-value is the indicator of how likely something is due to pure randomness, and so a low p-value means that something is very unlikely to be due to pure chance. For all of the variables that affect the instructors evaluation, the p-value is 7.822e-15, which is really small. This indicates that the variables present do in fact affect the evaluation of an instructor, and there is a small chance that it is due to randomness. The estimate shows how much each variable is able to affect the evaluation of an instructor as it increases. The p-value for each again indicates how likely it is that each variable could be affecting the evaluation due to pure chance. Based on the data I use the 95% rule to say that there is less than 5% that the strong affecting variables are not due to chance and they have a great impact, these include: minority, gender, credits, beauty, native, students, allstudents.

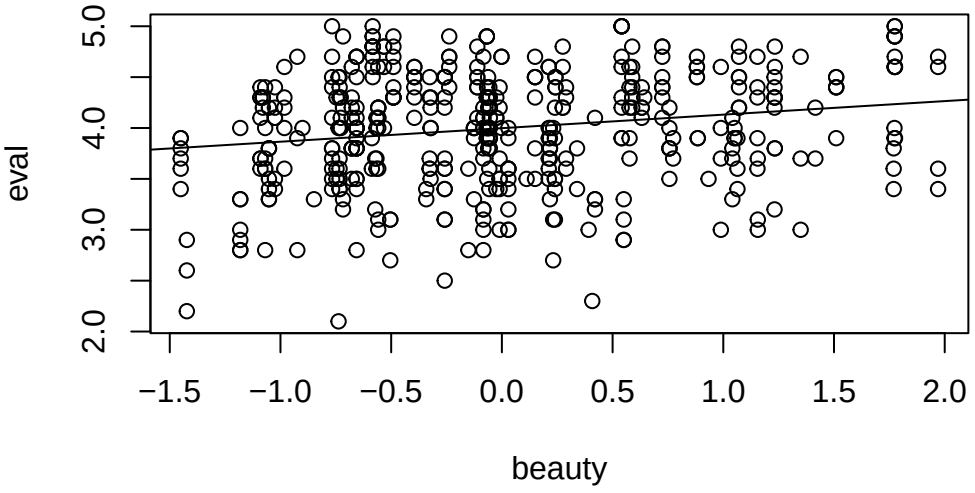

Figure 4: DON'T FORGET

Coefficient	Estimate	Standard Error
(Intercept)	3.80147758	2.185771339
minorityyes	-0.16522174	0.076515480
age	-0.00205587	9 0.002676510
gendermale	0.20347771	7 0.051443208
creditssingle	0.56053261	6 0.116679935
beauty	0.14575801	5 0.032194375
divisionupper	-0.04479519	3 0.056791711
nativeyes	0.23569123	3 0.106458510
tenureyes	-0.05123346	6 0.062578312

Coefficient	Estimate	Standard Error
students	0.007426996	0.002300905
allstudents	-0.004626833	0.001396391

Table (**strong-coefficients-affecting-eval?**) shows all of the variables that affect an instructor's evaluation the most. The estimate is how much the evaluation is how much the evaluation will either increase or decrease based on the change in the variable, and the p-value is how likely it is to be due to randomness.

Table 3: DON'T FORGET

Coefficient	Estimate	Standard Error
(Intercept)	3.801477582	0.185771339
minorityyes	-0.165221740	0.076515480
age	-0.002055879	0.002676510
gendermale	0.203477717	0.051443208
creditssingle	0.560532616	0.116679935
beauty	0.145758015	0.032194375
divisionupper	-0.044795193	0.056791711
nativeyes	0.235691233	0.106458510
tenureyes	-0.051233466	0.062578312
students	0.007426996	0.002300905
allstudents	-0.004626833	0.001396391

4 Discussion

#4.1 Beauty and Evaluation After reviewing all of the data, my initial hypothesis that beauty does influence an instructor's evaluation stands true, as the data shows that for each unit increase in beauty should result in a corresponding increase in evaluation score of 0.16. However, other factors also play a significant factor to the instructors evaluation.

4.2 Male and Female Instructors

From the findings, it has been shown that female instructors receive lower evaluations than male instructors, while holding all other variables constant. The average female evaluation was 3.90, while the average male evaluation was 4.07, despite females scoring higher in terms of beauty than males. The difference is small, however statistically significant, as it contradicts my hypothesis of higher beauty leading to a higher evaluation. However, more data and research would need to be done in order to explore the possible reasons, and see if the trend continues.

4.3 Class Size

Another factor which affected the evaluation of instructors, was the class size. The data shows that instructors who teach larger classes receive higher evaluations, than those who teach smaller ones. However, the following trend is seen to take less of an effect as class sizes continue to increase. As such, there is a positive correlation between the class size and evaluation, but this effect is limited. # 4.4 Minorities Instructors who are minorities are seen to receive lower evaluations than those who are not a minority. The data shows that minority instructors received evaluations that are 0.16 units lower on average than those instructors who are not a minority.

4.5 Native

Based on the data, instructors who are native receive lower evaluations than those who are not. The coefficient estimate for "nativeyes" is 0.259451, meaning that being a native instructor would increase the evaluation score by 0.259451 units, as compared to those who are non native. This effect has a significant effect on the evaluation of the instructors with a p-value of 0.013446. However, according to the data, non-native instructors actually receive higher evaluations than their native counterparts. This seems counter intuitive that while the estimate is positive, they receive lower evaluations, however, the estimate is holding all other variables constant which means that the estimate is the difference in the evaluation is everything else about two instructors remains identical. As such, though the estimate is positive, the actual

data shows that the native instructors actually receive lower evaluations that the non-native instructors.

4.1 Limitations

There are several limitations to consider from the following study. Firstly the study was conducted at one university, and therefore cannot be a generalization of other universities or students. Additionally, the study is relatively, old given by the fact that it was conducted in the years 2000-20002, and cannot reflect the changes in student demographics, teaching methodologies, and evaluation practices since then. Additionally, the method of data collection is unknown, therefore the reliability of the data is unknown. Moreover, the grading scale which was used in the study ranges from -2.0 to 2.0 for measuring beauty, which could potentially be unfamiliar to some and cause invalid ratings. The measure of beauty in the study is also based on personal preference of the students which were assessing the instructors beauty, which cannot be a unified measurement, and would vary depending on which students were chosen to evaluate the instructors beauty. This could potentially introduce variability and bias in the study.

Going back to the results of the native instructors versus the non-native ones. While the "native-yes" estimate is positive, indicating that native instructors should receive a higher evaluation, the actual data shows they receive a lower one. This means there could be some hidden variables which have not been accounted for such as the instructors language proficiency, that may also affect the students evaluation of their instructors.

Lastly, the study lacks information about the students themselves which carried out the evaluations. Cross referencing instructor evaluations to student data such as their age, gender, academic performance could have significant impact on the instructor. For example female students could find a male instructor highly attractive, but if only 6 male students were chosen to evaluate their beauty, then the data could be skewed. Additionally, students which do well in a course are more likely to review their instructor positively, while those who performed poorly tend to leave more negative evaluations. Leaving an abnormally negative score, could lead to the results being inaccurate, and without said information it is hard to understand the factors which influence an instructor's evaluation.

Next Steps

- would want to change the study collection
-

Kleiber, Christian, and Achim Zeileis. 2020. *Applied Econometrics with r.* https://cran.r-project.org/package=AER.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Team, RStudio. 2021. *RStudio: Integrated Development Environment for r*. Boston, MA: RStudio, PBC. https://www.rstudio.com.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. *Welcome to the tidyverse*. *Journal of Open Source Software*. Vol. 4. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*.

Wickham, Hadley, and RStudio. 2021. *Testthat: Get Started with Testing*. https://testthat.r-lib.org/.