STAT 151A Final Project

Jacob Sharadin (3036925877)
Peter Shen (3039143525)
Maggie Vashel (3040689798)

Do easy graders get better evaluations?

In this paper we aim to address the question of whether instructors that give higher grades receive better evaluations from their students. This question is important because if students give better evaluations to instructors that grade more leniently, then there can be an inherent incentive for teachers to grade easier if they want to have better evaluations, either for their own career benefit (for instance, towards their own evaluation for tenure), for their reputation, or their own sense of accomplishment as a teacher. This can contribute to a trend in U. S. higher education towards grade inflation (Carter & Lara, 2016). Given evidence for such an effect, when faculty are evaluated based on their student feedback, these evaluations can potentially be adjusted based on grading, or taken in the context of the grading standards of that instructor.

Source Data

To address this question, we use a dataset of student evaluations from Owen et al, 2024, a study on gender bias in student evaluations. The dataset contains 2,719 observations with 19 variables from a selective U.S. liberal arts college. Our dependent variable is positivity, which is a rating from 0 to 5 of the positivity of the student's evaluation, derived from the average of the positivity rating of three reviewers. Our independent variable of interest is facultygrades, which is the average of (grade/term GPA) for all students in the class (where grade/term GPA is the student's grade in that class divided by their GPA for the semester). This is a measure of how high or low the teacher's overall grading is, since it compares the average grade in their class to the average grades across all classes in the sample.

We include additional variables to control for potential confounds stemming from demographic factors, factors related to the characteristics of the classes, and teacher characteristics – in several versions of the model the exact regressors included vary.

For demographic factors:

- Pellgrant - whether students have received a Pell grant (a proxy for socioeconomic status since this is a need-based grant)
- Whitestudent - whether students are white (race/ethnicity)
- Female - whether students are female (gender)

For course related factors:

- Science/socsci - whether the class is in the social sciences or sciences
- Class_size - quartile class size (1 to 4)

For teacher characteristics

- Yrs_experience - number of years of teaching experience (bucketed from 1 to 3)

- Femalefaculty - gender of teacher

Additionally, we use student ID as a random variable to control for individual differences in evaluating professors, since there are some students that appear in the datasets multiple times (in multiple classes). Given the repetition, the data are not independent.

The data were collected in a random control trial for the study around gender bias. These study groups aren't relevant for our analysis, but we include the group number as a potential confound to control for as one of the groups was informed of their final grade before submitting their evaluation.

We also include several interaction terms:

- Female*science/socsci - in case there is an effect of female students in the sciences (vs the social sciences) as a potentially underrepresented group.
- Femate*femalefaculty - in alignment with the original study, controlling for any effect of student gender relative to instructor gender.

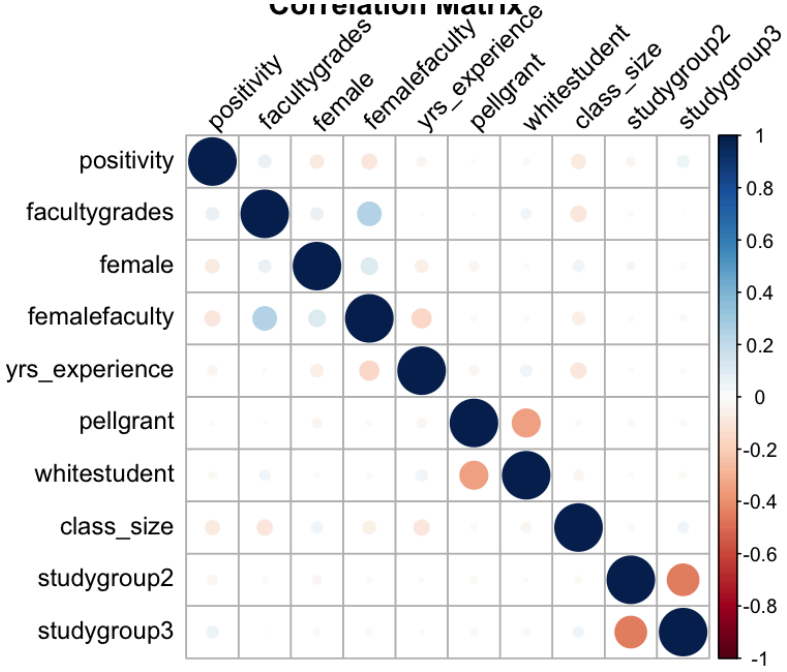
While this data has a relatively large number of records and includes many relevant confounders in addition to our DV and IV of interest, there can be some potential shortcomings. Significantly, we don't know the actual quality of the instructor's teaching, presumably an important predictor of evaluations.

EDA and Data Cleaning

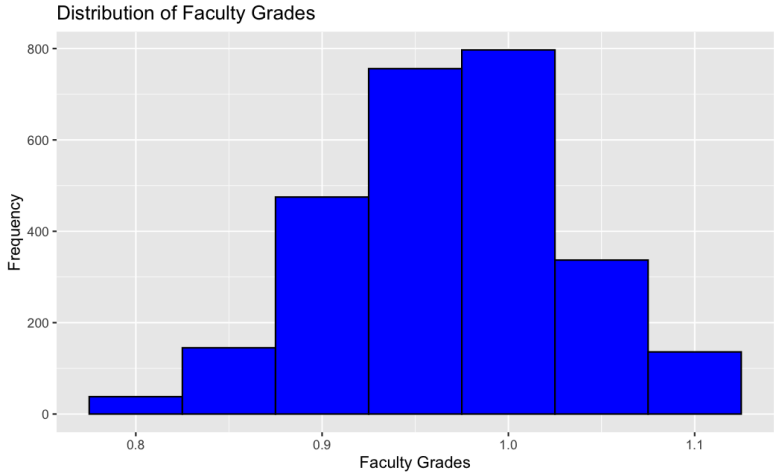
Missing Values - We checked for missing values in our dataset using a column-wise analysis. Missing data was identified in variables such as facultygrades, class_size, and positivity. These observations with missing values were removed to ensure consistent and unbiased analysis. For example, if facultygrades is missing for professors who grade leniently, we may underestimate the effect of grading leniency on evaluations. This was also done to reduce the risk of introducing noise or error due to incomplete data. In total, 440 observations were omitted.

Outliers - We identified outliers using leverage scores, influence measured by Cook's distance, and residual diagnostics from our initial regression model. Outliers can distort parameter estimates and lead to unreliable conclusions. For example, a professor with extremely high grades might disproportionately influence the effect of facultygrades on positivity. As part of our data cleaning, we identified all high-leverage/influence observations and removed these 163 outliers to meet regression assumptions, such as normality and homoscedasticity, by minimizing the impact of outliers on residuals. After excluding outlier observations, the regression model was refitted on the cleaned dataset to evaluate changes in parameter estimates and model performance.

Correlation Matrix - The correlation matrix provides an overview of the relationships between key variables in the dataset. Strong correlations are indicated by darker colors, while weak or negligible correlations are shown with lighter colors. Most correlations are relatively weak, suggesting minimal linear dependencies among variables.



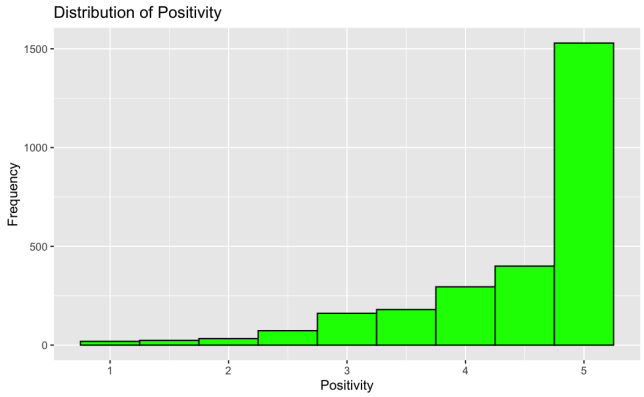
Distribution of Faculty Grades - The histogram of facultygrades shows a relatively normal distribution centered around a mean close to 1. This suggests that the grading behavior of professors is relatively consistent, with few extreme values.



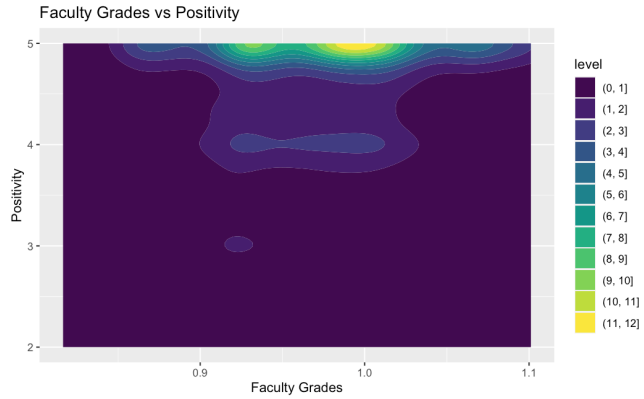
Initial Regression - Positivity ~ Faculty Grades As part of EDA, we did an initial check on the relationship between facultygrades and positivity in a simple regression model with just the one regressor and found a significant effect:

```
Linear Model Summary (model
================================================
                        Dependent variable:
                    ---------------------------
                               positivity
------------------------------------------------
facultygrades                   0.678***
                                p = 0.003

Constant                        3.906***
                                p = 0.000

------------------------------------------------
Observations                      2,244
R2                                0.004
Adjusted R2                       0.004
Residual Std. Error         0.646 (df = 2242)
F Statistic             8.929*** (df = 1; 2242)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

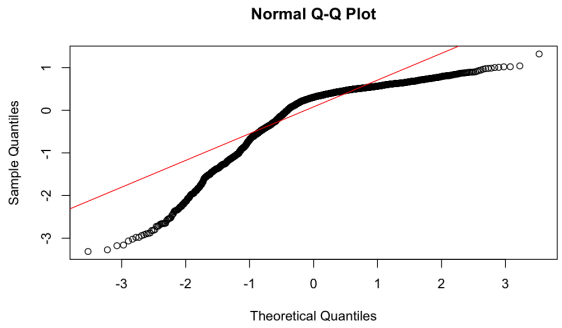
Distribution of Positivity - The histogram of positivity is highly right-skewed, with the majority of responses concentrated at the maximum value. Despite this, after careful consideration we opted not to transform the variable to increase interpretability in this inference model.



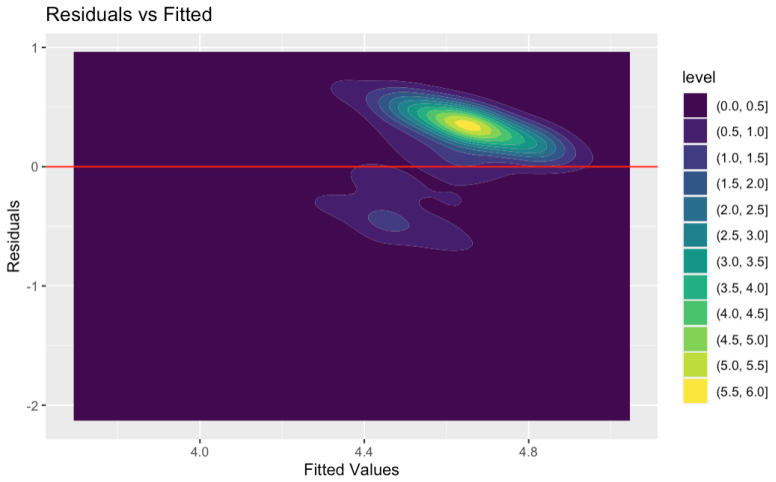
Linearity - A scatterplot of facultygrades vs. positivity showed a generally linear spread, although the discrete nature of the IV, facultygrades, limited the appearance of smooth linearity.



Normality of Residuals - A Q-Q plot of residuals revealed deviations from the expected straight line at the tails, indicating potential non-normality.



Homoscedasticity - A plot of residuals vs. fitted values revealed heteroscedasticity. Residuals show a slight wider spread at lower fitted values, and a systemic downward sloping. To address this, robust standard errors (using the sandwich estimator) were used in subsequent analyses.



Analysis

Independence - Since there are some students that appear in the datasets multiple times (in multiple classes), this hierarchical distribution suggests that independence is not met. We used a mixed model approach to account for this, by using student ID as a random variable to control for individual differences in evaluating professors.

Full Model - This is the full model we've used to capture the effect of faculty grades on student evaluations. We used additional small and medium models with some subset of these variables to further inform our inference.

$$\begin{aligned}\text{positivity} = {} & \beta_0 + \beta_1 \cdot \text{facultygrades} + \beta_2 \cdot \text{female} + \beta_3 \cdot \text{femalefaculty} \\ & + \beta_4 \cdot \text{yrs\_experience} + \beta_5 \cdot \text{pellgrant} + \beta_6 \cdot \text{whitestudent} \\ & + \beta_7 \cdot \text{class\_size} + \beta_8 \cdot \text{studygroup2} + \beta_9 \cdot \text{studygroup3} \\ & + \beta_{10} \cdot (\text{female} \cdot \text{science}) + \beta_{11} \cdot (\text{female} \cdot \text{femalefaculty}) \\ & + u + \epsilon\end{aligned}$$

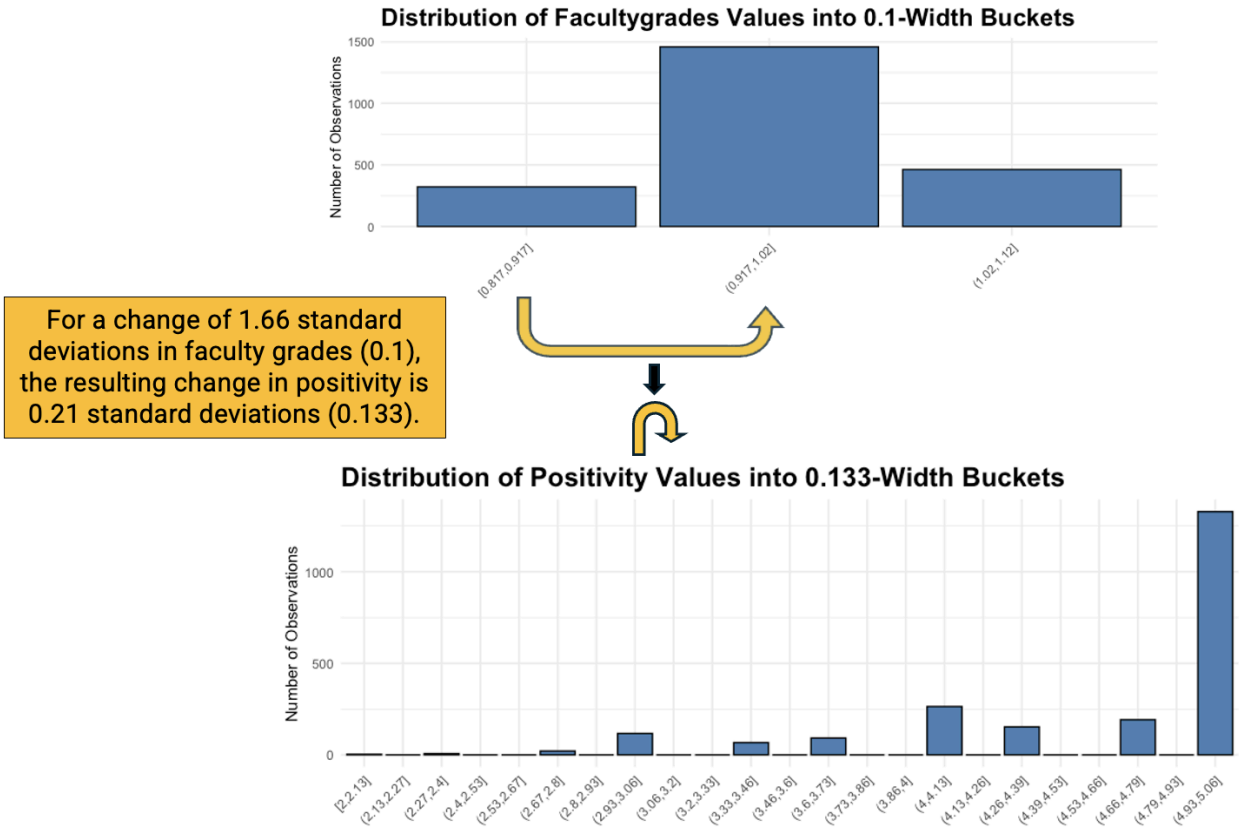
Where u is our student random effect and e is the error term

The variable of interest is facultygrades. Since our regression includes a random effects term we'll run "lmer" instead of "lm" in R. "lmer" incorporates random effects to account for variations due to individual students, since some students take multiple classes. The inclusion of random effects can improve the robustness of our estimates.

	Dependent variable:				
	positivity				
	OLS				linear mixed-effects
	(1)	(2)	(3)	(4)	(5)
facultygrades	0.935***	1.029***	1.029***	1.590***	1.331***
	p = 0.002	p = 0.00001	p = 0.00001	p = 0.00002	p = 0.00001
treated			0.065**	0.027	0.071***
			p = 0.018	p = 0.442	p = 0.009
female			-0.111***	-0.139**	-0.140***
			p = 0.0001	p = 0.035	p = 0.007
femalefaculty		-0.172***	-0.170***	-0.187***	-0.152***
		p = 0.000	p = 0.000	p = 0.003	p = 0.002
yrs_experience		-0.030*	-0.039**	-0.058***	-0.044***
		p = 0.053	p = 0.013	p = 0.005	p = 0.005
pellgrant			0.005	-0.039	-0.0003
			p = 0.898	p = 0.400	p = 0.994
whitestudent			-0.051*	-0.040	-0.048
			p = 0.089	p = 0.312	p = 0.138
class_size			-0.050***	-0.072***	-0.063***
			p = 0.00005	p = 0.00001	p = 0.00000
science				0.019	0.007
				p = 0.797	p = 0.908
socsci				0.081	0.095**
				p = 0.121	p = 0.017
female:science				0.087	0.101*
				p = 0.261	p = 0.096
female:femalefaculty				-0.014	-0.011
				p = 0.852	p = 0.843
Constant	3.521***	3.732***	3.942***	3.367***	3.652***
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Observations	2,376	2,236	2,236	2,376	2,236
R2	0.004	0.021	0.040	0.034	
Adjusted R2	0.004	0.019	0.036	0.029	
Log Likelihood					-2,138.078
Akaike Inf. Crit.					4,306.156
Bayesian Inf. Crit.					4,391.843
Residual Std. Error	0.855 (df = 2374)	0.633 (df = 2232)	0.628 (df = 2227)	0.844 (df = 2363)	
F Statistic	10.616*** (df = 1; 2374)	15.721*** (df = 3; 2232)	11.477*** (df = 8; 2227)	6.994*** (df = 12; 2363)	
Note:					*p<0.1; **p<0.05; ***p<0.01

The final model, model #5 above, gives a coefficient estimate of 1.331 for facultygrades, with a 99% confidence interval of [0.612, 2.050] (robust SE, using clubSandwich estimator in R) and standard error of 0.278.

Interpretation - The coefficient estimate from our model suggests that for every one-unit increase in facultygrades, positivity rating is expected to increase by 1.331 points on the 1–5 scale, holding all other variables constant. As shown in the visual below, given the range and distribution of each variable, this is a very small effect size, most likely not impactful in the real life interpretation of student evaluations. For a change of 1.66 standard deviations in faculty grades, the resulting change in positivity is 0.21 standard deviations.



Discussion/Conclusion

In conclusion, the grades that faculty give (or the impression students have of the grades they will receive in the future) influence evaluations significantly, but with a very small effect size, such that instructors get (slightly) more positive evaluations when they grade higher on average.

Based on these findings – and the pervasiveness of grade inflation – future research into the relationship between evaluations and grading practices is warranted. However, we would not recommend changes to university practices at this time given that the effect size is not large enough to be meaningful.

In this analysis there are limitations, and assumptions have been made:

- Residuals violated the normality assumption, and we opted not to transform the skewed response variable positivity for reasons of interpretability. This is hopefully at least partially corrected for with the robust “sandwich” covariance calculation for the CI.
- We assume that "Positivity" as rated in this study data reflects positive evaluation of teachers by students.
- The likelihood that other confounding variables exist that are not in this dataset is a limitation. Importantly, we don't have an objective measure of how good the teacher

actually is. Additional confounders not in the dataset could include how interesting the class is and the quality of course materials.

- There is a potentially problematic circularity, because It may be the case that the students of an excellent teacher objectively do better, and thus deserve better grades, depending on the assumptions about grading standards (for instance, should a curve apply regardless of the overall quality of the student work).
- These data are a limited (although reasonably large) sample from one school and several departments. We are assuming that these data are representative of other classes, students, and schools. The findings need to be validated across a broader sampling of universities, courses, and instructors.
- Students in groups 1 & 2 don't actually know what their final grade is at the time of evaluation. We are assuming that they have a sense of their grade that is accurate - or we are actually assessing student's impression of how the faculty grade rather than the actual final grades.

Despite these limitations and assumptions, we conclude that these data support the conclusion that faculty grading standards do have a significant effect on student evaluations.

References

Carter, M., & Lara, P. (2016). Grade inflation in higher education: Is the end in sight?. *Academic Questions*, *29*(3).

Ann L. Owen, Erica De Bruin & Stephen Wu (02 Oct 2024): Can you mitigate gender bias in student evaluations of teaching? Evaluating alternative methods of soliciting feedback, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2024.2407927