Categorical Data Analysis: SOTE Investigation

Joey Fitch

Nate Kotila

Jeff Lee

San Jose State University

December 7, 2017

Contents

1 Introduction

Every semester, instructors at San Jose State University receive "Student Opinion of Teaching Effectiveness" surveys, or SOTE's. Here, the students provide ratings as responses to questions that are formulated for the purpose of assessing the instructors. These ratings influence tenure and promotion of the instructors. In this study, we are interested in exploring student ratings and whether or not certain external characteristics influence the ratings. Our aim is to identify whether instructor ratings are affected by factors other than true teaching effectiveness.

2 The Data

Our data is obtained from a cross-sectional study of 169,429 SOTE responses, from the Fall 2014 and Spring 2015 semesters. The data provides the student's standing (freshman, sophomore, etc.) and then a bunch of information on the classes, such as course subject, number of enrolled students, the enrollment cap, and type of instruction. SOTE survey responses are recorded for the following: 12 questions for students to evaluate certain aspects of instructor effectiveness, and a 13th question for overall instructor effectiveness. All variables are discrete; the enrollment counts are ratio measures, but all other variables are nominal, ordinal, or interval measures. Specifically, we will consider grades to be a interval measure, as grades measure on a standard 10% scale per University Standards (each grade corresponds to the same 10% range of percentage grades). A more detailed description of the data can be found in Appendix A.

3 Data Processing

To effectively investigate the association between "perceived teaching effectiveness" and the other student characteristics, we process the data in a number of meaningful ways.

3.1 Aggregating the Teacher Ratings

Our data contains 13 questions related to instructor effectiveness ratings. The thirteenth question is an overall rating, aggregating ALL of the rating questions. If this aggregate score is truly an adequate representation of the previous 12 questions, then we could narrow our analysis to this one single response variable. This would allow us to circumvent the potential problem of highly correlated covariates and avoid the inflated error rates from looking at each question individually.

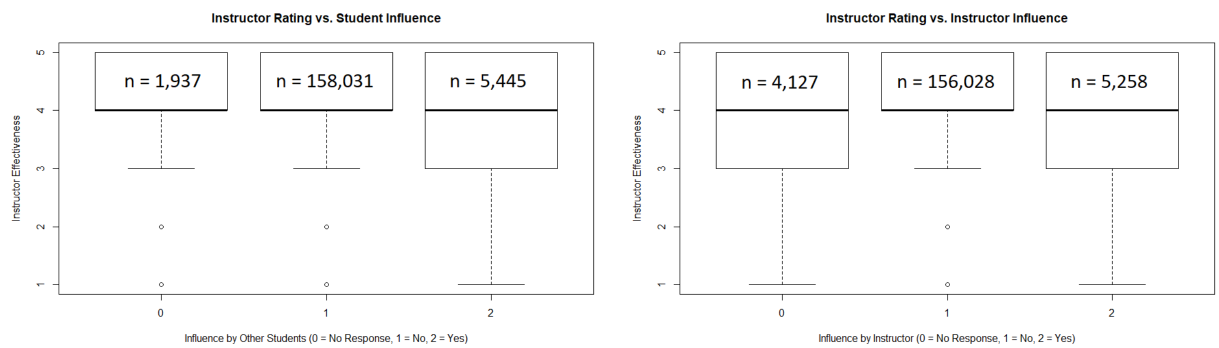
To determine the validity, we explore the association between each of the first twelve "specificity" questions (asking about individual aspects of the professor's teaching quality) and the aggregate thirteenth question ("Overall, this instructor's teaching was..."). Every specificity question demonstrates a high concordance proportion with the aggregate thirteenth question:

Question:	1	2	3	4	5	6	7	8	9	10	11	12
Concordance with Q13:	0.95	0.95	0.96	0.95	0.96	0.94	0.93	0.95	0.94	0.94	0.97	0.95

The high concordance indicates that question 13 can effectively represent the majority of information contained in the first 12 questions. We thereby verify the usage of question 13 as a strong overall measure of teaching effectiveness, avoiding the problems related to individual assessment of each specificity question.

3.2 Biased Observations

We immediately identify a potential source of bias in the data, where some students report "undue influence from the instructor [and/or] other students" in their survey responses. This essentially provides a self-reported indicator of bias, casting doubt on the validity of all other self-reported responses in the questionnaire.We construct boxplots to check for any explanation of this response:



The vertical axes show the distribution of instructor ratings within each response. It appears that respondents who reported "undue influence" have a higher tendency to rate instructor effectiveness = 3, though deeper inspection of the data reveals that this is not a strong distinction: 77% of "0/1" respondents rate 4 or higher, while 72% of "2" respondents give the same rating. This value is coincidentally near 75%, skewing the discrete-valued quartile in the boxplot. There is not enough evidence to say anything truly substantial about these values, so – to avoid contaminating our sample with biased values – we exclude these 12,376 observations from our data.

3.3 Missing and Invalid Values

When investigating the association between teacher effectiveness and student grades, the data contains certain values that are non-usable in answering this question. These values are either omitted by the student, or they do not give adequate information about student performance that can be used in our analysis. For example, question 14 (expected grade) has a response labeled "Other" which indicates that the student is expecting to receive either credit, no credit, have an incomplete, or – in the dataset's own words – "et cetera". This single option encompasses students who are passing, failing, or not even finishing the semester, so it is impossible to categorize these students in a meaningful way. As such, we discard any observations with an "Other" response from our data. We do the same for missing responses, as our entire analysis requires knowledge of this value.

Likewise, we omit any students whose real grade is "credit", "no credit", "withdrawal", "withdrawal unauthorized", "incomplete", "report delayed", or "report in progress". We cannot quantify these students' performance in a meaningful way, relative to the vast majority of students who report a grade according to the standard A-through-F scale.

We also note that there is a group of 79 observations that are categorized as "Undeclared" for college membership. Since this is very small relative to the rest of the other groups of colleges (which all have over 9000 observations), we assert that the Undeclared group may not represent the general target population of students, and is too small to draw meaningful conclusions about. Therefore, we exclude the Undeclared observations from our data analysis.

In total, we drop 10,737 observations during this data cleaning, constituting 14.6% of the original data (not including the biased observations). This should not reduce much certainty from our conclusions.

3.4 Discrepancy of Expected Grade vs. Actual Grade

We investigate an interesting quantity in the data, measuring the difference between a student's expected grade and the grade they actually receive. As discussed in section 2, the grades correspond to an interval measurement scale, so subtraction is valid. We define "grade discrepancy" to equal (expected grade – actual grade), where {A, B, C, D/F} correspond to {4, 3, 2, 1}, respectively. Our data shows the following distribution for grade discrepancy:

Discrepancy	–3	–2	–1	0	1	2	3
Frequency	369	3,198	28,219	95,157	23,658	1,201	32

where a negative discrepancy indicates optimism: "low actual grade vs. high expected grade", and a positive discrepancy indicates pessimism: "high actual grade vs. low expected grade". Discrepancy = 0 indicates that the student received the grade that they expected. We see that 63% of students accurately estimated their grade. Of the students with inaccurate grade estimates, we see a larger tendency for optimism (overestimating their actual grade, negative discrepancy) than for pessimism (underestimating their actual grade, positive discrepancy).

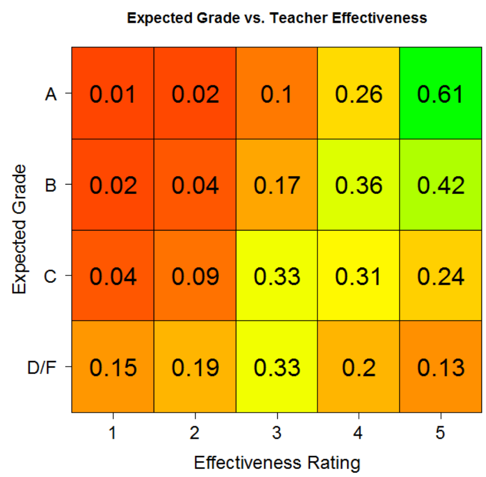
It is worth noting that *any* student can have 0 discrepancy, while only "A" students can have discrepancy = 3 (since the only way to achieve discrepancy = 3 is to expect an F but receive an A). Likewise, only "D" or "F" students can have discrepancy = –3 (since this is only possible by expecting an A but receiving a D or F). Similar restrictions hold for any non-zero discrepancy value.

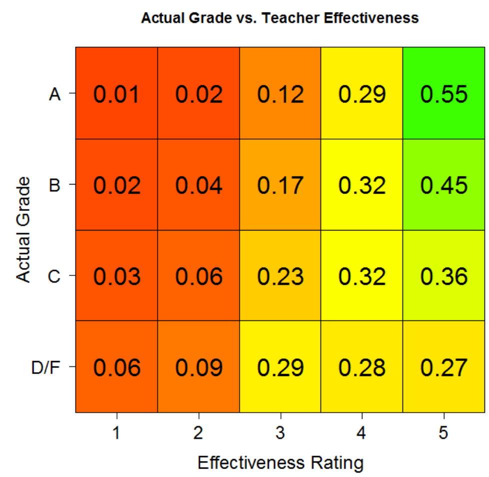
4 Exploration of the Data

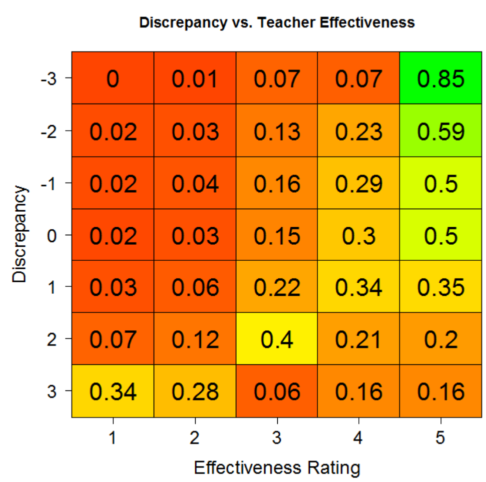
We lead some exploratory analysis of the data, to detect any patterns of association between variables. We are particularly interested in whether any variable can predict a student's aggregate rating of teacher effectiveness (question 13).

4.1 Expected Grade and Actual Grade vs. Teacher Rating

We are interested in the conditional distribution of teacher effectiveness ratings, conditioning on students' expected grades, actual grades, and/or grade discrepancy. To determine whether any associations exist, we construct three separate frequency tables – measuring each of the three grade measures vs. teacher rating – and estimate the conditional probabilities for each row.







Each cell represents the conditional probability of a specific effectiveness rating, conditioning on the row value (expected grade, actual grade, or discrepancy). Green indicates high probability, while red indicates low probability.

Recall that negative discrepancy indicates optimism, where students overestimate their actual grade, and positive discrepancy indicates pessimism, where students underestimate their actual grade.

We immediately see some striking patterns. Both expected grade and actual grade demonstrate strong association with the effectiveness rating. Expected grade appears to show a sharper association: of students expecting to fail, 34% assigned an effectiveness rating of 1 or 2, while only 15% rated a 4 or 5. In contrast, for students expecting an A, only 3% rated a 1 or 2, while 87% gave a "5" rating. In general, students who expect to fail appear more willing to assign low effectiveness ratings. The conditional distributions for the actual grade show a similar pattern, although the disparity appears somewhat less stark – expected grade appears to be a stronger predictor for teacher evaluations.

Even beyond their individual values, the discrepancy between expected grade and actual grade appears to be the most informative metric. We see a very stark difference between optimistic discrepancy and pessimistic discrepancy. The optimists with negative discrepancy (expected grades higher than reality) give a high effectiveness rating, most of the time. As discrepancy increases (moving deeper toward pessimism), we see that the distribution of ratings skews far stronger toward low effectiveness, relative to other students' ratings.

For our model, we will use grade discrepancy to be one of our variables. Even though grade discrepancy is a linear combination of real grade and expected grades, if we only use two instead of all three, we will not run into problems later on in statistical analysis, especially since grade discrepancy is not necessarily strongly correlated with expected grades. From our data, the estimated correlation between grade discrepancy and expected grade is -0.30 and the concordance proportion is 0.28. With this in mind, we also will use expected grades as a variable in our model.

4.2 Student Demographics vs. Teacher Rating

Beyond expected and actual grades, the data offers some demographic information about the students which may help further explain some variability in the teacher evaluations.

It is reasonable to believe that "subject of study" may be of interest, as teachers from different disciplines may resonate differently with students. For example, students may find a business professor to be more dry or ineffective than a professor of statistics, whom are widely adored by their students. However, the data contains 121 different course subjects, which poses some issues. The sample sizes for each subject vary tremendously, ranging from 2 observations all the way up to 6211 observations. 28 of the 121 subjects have less than 100 observations total. The small sample sizes obstruct meaningful analysis, as our estimates would carry a large amount of uncertainty. Moreover, making hundreds of comparisons across different subject levels would require many parameter estimations and massively inflate our error rates, presenting another obstruction of meaningful analysis.

To mitigate these concerns, we instead consider the *college* variable as a meaningful grouping of similar course subjects. There are 8 distinct college values, providing a reasonable number of partitions, and the 7 main colleges have roughly 10,000 observations or more (the "undergrad" category has only 79 observations). We observe the frequency table for college vs. instructor effectiveness rating:

College	Instructor Rating 1	2	3	4	5	Sample size
Applied Sciences & Arts	0.02	0.03	0.14	0.29	0.53	26717
Grad. School of Business	0.02	0.04	0.19	0.31	0.44	21398
Education	0.01	0.03	0.14	0.30	0.51	9794
Engineering	0.03	0.06	0.21	0.33	0.37	23755
Humanities	0.01	0.03	0.13	0.30	0.52	21766
Science	0.03	0.05	0.19	0.30	0.43	20619
Social Sciences	0.01	0.03	0.14	0.30	0.52	27630

We see an interesting pattern here. The distributions of the 1, 2, and 4 ratings are relatively even between the various colleges. However, we see a direct exchange between the 3 and 5 ratings: wherever there is a low amount of 3 ratings, we see a larger proportion of 5 ratings (and vice versa). The large sample sizes allow a high degree of certainty in those estimates as well, so a spurious association is unlikely.

In order to apply any standard measure of association, we need to collapse the 5 effectiveness values into a set of 2. The natural grouping would be "positive ratings" versus "negative ratings". Ratings of 1 and 2 are obviously negative, while 4 and 5 are conversely positive. We choose to define 3 as a negative rating because, as students ourselves, we believe that most students would only assign a 3 rating if they actually have problems with an instructor's teaching. The majority of students with no qualms would likely assign a "4" or "5" even if the instructor is not excessively great, as long as the student has no major complaint. This indicates 3 as a relatively negative rating. The opposition between ratings of 3 and 5 shown in our data supports our splitting decision.

Collapsing the table in this way yields the following frequency table:

College	Applied Sciences & Arts	Graduate School of Business	Educ.	Engin.	Human. & Arts	Science	Social Sciences
Pr(positive)	0.82	0.75	0.81	0.70	0.82	0.73	0.82
Sample Size	26,717	21,398	9,794	23,755	21,766	20,619	27,630

We define the new variable "positive" to be an instructor rating of 4 or 5. Conversely, "negative" constitute ratings of 3, 2, or 1. This table shows that college does seem to influence teacher rating, which supports our interest of the college variable, and so we will proceed with the usage of this variable later on in statistical analysis.

We also considered class level as a potential predictor of student's SOTE ratings (e.g. freshman may behave different from graduate students), but upon investigation of teacher ratings versus student levels, we did not find any meaningful association between student level and teacher ratings.

This, combined with the advantage of having a more parsimonious model, is why we will proceed without using class level in our statistical model.

5 Statistical Method and Interpretation

In our exploration of the data, we discussed the usage of grade discrepancy and expected grade as variables in our model. We also discussed the usage of college as a demographic variable. These three variables will be used predict our response variable, teacher ratings, in our data analysis.

The hope is that through this analysis, with contextual knowledge, we will be able to discuss about any outside influence on instructor evaluation beyond the instructors own teaching capabilities.

5.1 Three-Way Associations and Heterogeneity

We have shown through exploration and visualization that association of instructor evaluations with both college and student grades (expected grade and grade discrepancy) should not be ruled out. It is reasonable to question whether student grades have a different association within each college – business graduate students may have very different attitudes about their grades than humanities undergraduate students.

To investigate homogeneity (by conduncting the Breslow-Day test), we must collapse each variable to a binary measure (except for college). For instructor evaluations, we collapse the 1 through 5 scale down to two groups, 1:3 and 4:5 (as detailed in section 4.2). For expected grade, we group together the passing grades (A through C) against the failing grades (D and F). This is the most sensible distinction in the context of the problem, and the division between the groups is supported by the heatmap frequency tables of section 4.1. The failing grades show a substantially different distribution of evaluations than the other expected grades.

Discrepancy presents a more interesting decision. There are three qualitatively different situations: the optimists, the pessimists, and the realists. We postulate that the pessimists would probably give the most abnormal ratings: optimists would most likely tend to give high ratings, and we've shown that the realists tend to naturally skew toward high ratings anyway. The pessimists deviate from this trend by giving low ratings, distinguishing them as the "odd" group. Under this logic, we collapse the ratings to pessimists (positive discrepancy) versus realists & optimists (discrepancy ≤ 0). We support this notion with data in Appendix B.

We run two Breslow-Day tests for homogeneity at 95% significance, checking whether college has any influence on the association between SOTE ratings and student grades:

College vs. Evaluation vs. Expected Grade	College vs. Evaluation vs. Grade Discrepancy
$X^2 = 21.9$	$X^2 = 17.2$
$\chi^2_6(.95) = 12.6$	$\chi^2_6(.95) = 12.6$
$p = 0.0012$	$p = 0.0085$

This result suggests that our associations do change under different levels of the college variable. To investigate the specific relationships, we estimate the relative risk of positive evaluations for both variables, under each college factor:

Relative Risk		College 1	2	3	4	5	6	7
$\dfrac{\widehat{Pr}\left(\text{Positive Eval. } \mid \text{Exp. Pass}\right)}{\widehat{Pr}\left(\text{Positive Eval. } \mid \text{Exp. Fail}\right)}$	=	3.99	3.04	2.83	2.46	3.87	2.64	3.46
		(0.001)	(<0.001)	(0.005)	(0.001)	(0.001)	(<0.001)	(<0.001)
$\dfrac{\widehat{Pr}\left(\text{Positive Eval. } \mid \text{Non-pessimism}\right)}{\widehat{Pr}\left(\text{Positive Eval. } \mid \text{Pessimism}\right)}$	=	1.81	1.60	1.54	1.59	1.61	1.51	1.59
		(0.001)	(0.001)	(0.004)	(0.001)	(0.002)	(0.001)	(0.001)

1 = Applied Sciences & Arts
2 = Grad. Business
3 = Education
4 = Engineering
5 = Humanities & Arts
6 = Science
7 = Social Sciences

Each cell represents the estimated relative risk for positive evaluations within that college. For example, a student in an Applied Sciences & Arts course is 3.99 times more likely to issue a positive evaluation when they believe they are passing, versus when they are failing. The parenthetical values indicate the half–width for a 95% confidence interval on each relative risk; with our large sample sizes, we have a high degree of certainty in all measurements. Students in graduate business, education, and STEM courses appear less prone to associate their evaluations with their expected grades, as smaller relative risks appear in these colleges of Science and Engineering. We postulate that failing is more common or accepted in STEM courses, influencing those students less. Graduate students are potentially more mature, placing more blame on themselves and less on the teachers. Students in education development are perhaps more appreciative of the system and identify with their professors, leading to less impressionability from their grades. Conversely, students of Arts and Social Sciences are perhaps less tolerant of failing. If true, then the expectation of failure would place more stress on these students. This stress may cause students to associate their professors with more negative emotions.

On the other hand, there does not seem to be any heavy association between grade discrepancy and College. Interestingly, there is almost perfect concordance between the two relative risk measurements, but discrepancy offers far less variation in its range of values. The Breslow-Day test for homogeneity was likely over-powered from the massive sample sizes, causing a hypothesis rejection in terms of statistical significance but not of practical significance. We therefore proceed without recognition of any major interaction between discrepancy and college.

5.2 Logistic Regression

Let $y_i = \begin{cases} 1 & \text{if the } i\text{th survey rates a 4 or 5.} \\ 0 & \text{if the } i\text{th survey rates a 1, 2, or 3.} \end{cases}$ This suggests a binomial distribution on y_i. We have shown previously that there is significant association of y_i with predictor variables "college", "expected grade", and "discrepancy". We discussed the effectiveness of College as a predictor of y_i back in section 4.2. To analyze the grade variables as predictors, we observe the relationship of each one with y_i:

5.2.1 Model Building

	Expected Grade			
	D or F	C	B	A
$P(y_i = 1 \mid \text{Exp. Grade})$	0.33	0.55	0.77	0.87

	Grade Discrepancy						
	–3	–2	–1	0	1	2	3
$P(y_i = 1 \mid \text{Discrepancy})$	0.91	0.82	0.79	0.80	0.69	0.41	0.36

At each subsequent level of expected grade, the distribution of teacher evaluations increases by a fairly consistent amount. This indicates a somewhat linear trend between the two, so we treat expected grade as a numeric, interval-valued predictor. Discrepancy, however, does not seem linearly related with evaluations. We see fairly even distributions for discrepancies from –2 to 0 and from 2 to 3, but with substantial probability shifts between other discrepancy values. In some sense, discrepancy is a actually modeling a latent variable for optimist/pessimism, and we don't know whether the relationship between those two variables is any sort of interval measure. The unit increase in discrepancy may not be meaningful, especially when a single category of discrepancy could encompass students of different backgrounds (for example, both "F" and "B" students could report a discrepancy=–1). For these reasons, we do not consider discrepancy as an interval measure, but rather as an ordinal measure. For usage with logistic regression, we apply it as a nominal category.

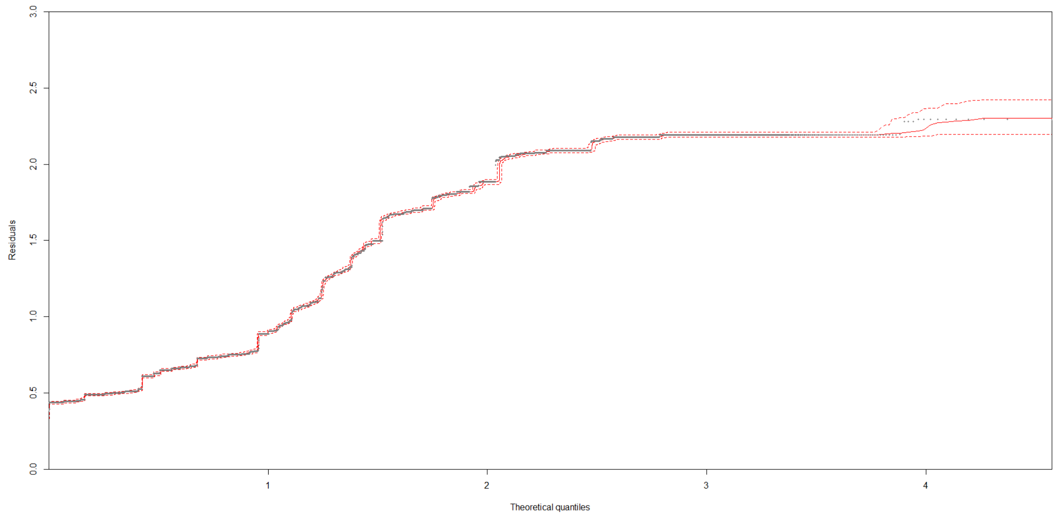
We explored the interaction between college and expected grade in section 5.1, determining that there was some three-way association with teacher evaluations. However, we already have 14 coefficients in our model, and interaction terms would blow this up greatly. To maintain interpretability of our model, we choose not to include the interaction terms in our model.

We model the parameter $\pi_X = E(Y \mid X)$ according to the logit link function:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \text{ Discrepancy (Factor) } + \text{ College (Factor) } + \text{ Expected Grade}$$

Because we have two categorical variables with 7 levels each, there are too many coefficients to write the full model equation here. The estimated parameters for our model can be found in Appendix D.

5.3 Model Diagnostics



Comparing the residuals of our model to a half-normal plot of residuals, it looks like our logistic regression model is a good fit for this data. Moreover, we check how our model does in prediction by training the model on a 90% subset and then measuring its accuracy on the 10% set of testing data. We see that it achieves 78.5% accuracy, but this comes with a caveat: it predicts a positive rating for nearly every response! This is to be expected though, as the substantial majority of observations are indeed positive. The model only predicts negative responses for observations with extreme positive discrepancy (=2 or =3), which is completely in line with the results of previous analyses – these are the only two subsets of data which produced a majority of negative teacher evaluations. Otherwise, the conditional expected value of y_i will always lean toward a positive rating.

6 Summary and Concluding Remarks

There is certainly more information roped into teacher evaluations than a pure assessment of teacher effectiveness. We saw that students' opinions of teachers are influenced by, at a minimum, the course subject and the students' own performance in the class. The large sample sizes allowed very little uncertainty in our conclusions.

For a true measure of teaching effectiveness, we recommend that instructors be compared only against other instructors within their same college. There is significant variability between colleges, which would not provide a genuine comparison of effectiveness between instructors. Furthermore, it may be most interesting to consider the grade discrepancies within each instructor's set of students. Clearly students with large discrepancies are very unsatisfied. If an instructor has a disproportionately large amount of students with high grade discrepancies, then that professor is not communicating effectively with his or students – this may be cause for concern.

A Details about the Data

Column Heading	Column Description	Code
Class Number	Class Number (from Catalog)	see http://info.sjsu.edu/cgi-bin/pdfserv?atok=catalog
Semester	Semester ID	2144 = Fall 2014; 2152 = Spring 2015
CRSID	Short Course Description	
Official Grade	Official Grade in Course	see http://www.sjsu.edu/registrar/students/grades-grades_changes/grade_symbols_and_values/index.html
Grading Basis	Type of Grades Offered in Course	CBE = Credit by Exam; CNC = Credit/No Credit; GRD = Grade
Level	Student Standing	1.LD = Lower Division; 2.UD = Upper Division; 3.GRAD = Graduate; 9.UNK = Unknown
Subject	Department Abbreviation	see http://www.sjsu.edu/academics/colleges_departments/
College	College Abbreviation	see http://www.sjsu.edu/academics/colleges_departments/
Student Level	Student Year in School	a.Fres = Freshman; b.Soph = Sophmore; c.Juni = Junior; d.Seni = Senior; e.2nd/ = Post-Bac; f.Cred = Credential Student; g.Grad = Graduate Student
Component	Type of Course Format	ACT = Activity; LAB = Laboratory; LEC = Lecture; PRA = Practicum; SEM = Seminar; SUP = Supervision
Total Enrollment	Number of Students Enrolled in Course	
Enrollment Cap	Enrollment Cap (from Catalog)	
Registered Count	Number of Students Registered for Course	
Open U Count	Number of Open University Students Enrolled	
Intruction Mode	Type of Instruction Format	M = Mixed; P = In-Person; PW = In-Person with Web Component; WW = Online; SW = Flipped (?)
Instrument	SOTE or SOLATE	
Record Number	Unique ID	
Question 1	*The instructor demonstrated relevance of the course content.*	0 = No Response; 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree
Question 2	*The instructor used assignments that enhanced learning.*	0 = No Response; 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree
Question 3	*The instructor summarized/emphasized important points.*	0 = No Response; 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree
Question 4	*The instructor was responsive to questions and comments from students.*	0 = No Response; 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree
Question 5	*The instructor established an atmosphere that facilitated learning.*	0 = No Response; 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree
Question 6	*The instructor was approachable for assistance.*	0 = No Response; 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree
Question 7	*The instructor was responsive to the diversity of students in this class.*	0 = No Response; 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree
Question 8	*The instructor showed a strong interest in teaching this class.*	0 = No Response; 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree
Question 9	*The instructor used intellectually challenging teaching methods.*	0 = No Response; 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree
Question 10	*The instructor used fair grading methods.*	0 = No Response; 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree
Question 11	*The instructor helped students analyze complex/abstract ideas.*	0 = No Response; 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree
Question 12	*The instructor provided meaningful feedback about student work.*	0 = No Response; 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree
Question 13	*Overall, this intructor's teaching was...*	0 = No Response; 1 = Very Ineffective; 2 = Ineffective; 3 = Somewhat Effective; 4 = Effective; 5 = Very Effective; 6 = N/A
Question 14	*What is your current estimate of your expected overall grade in this course?*	0 = No Response; 1 = A; 2 = B; 3 = C; 4 = D or F; 5 = Other (Credit/No Credit, Incomplete, etc)
Question 15	*You are a (freshman, sophmore, etc)...*	0 = No Response; 1 = Freshman; 2 = Sophomore; 3 = Junior; 4 = Senior; 5 = Graduate Student; 6 = Credential Only; 7 = Other (e.g., Open University)
Question 16	*Did you complete this form without undue influence from other students?*	0 = No Response; 1 = Yes; 2 = No
Question 17	*Did you complete this form without undue influence from the instructor?*	0 = No Response; 1 = Yes; 2 = No

B Discrepancy

Beyond plain intuition, the data justifies the grouping of 0 discrepancy with negative discrepancy (against positive discrepancy). We investigated the measure of association between these three types of discrepancy, calculating the relative risk for positive teacher evaluations:

0 Discrepancy vs. Positive Discrepancy		
Discrepancy	Good	Bad
0	73689	18146
Positive	16229	7707

With a relative risk of 1.18.

0 Discrepancy vs. Negative Discrepancy		
Discrepancy	Good	Bad
0	73689	18146
Negative	24213	6332

With a relative risk of 1.01.

We see that 0 discrepancy is very closely related with negative discrepancy, so this is the logical grouping of values when we collapse the Discrepancy variable down to 2 values.

C Logistic Regression Model

$$\log\left(\frac{\pi}{1-\pi}\right) = -1.079 + \text{Discrepancy} + \text{College} + \text{Expected Grade}$$

Discrepancy and College are treated as categorical variables. The tables below show the estimated values for the levels of each. Expected grade is treated as numeric, and can take the values 1 to 4, and a coefficient standard error of 0.009912.

Discrepancy		
Level	Value	Std. Error
-3	0	0
-2	-0.37107	0.19460
-1	-0.36463	0.18930
0	-0.29932	0.18890
1	-0.35055	0.18957
2	-0.83149	0.19924
3	-0.28811	0.44086

College		
Level	Value	Std. Error
Applied	0	0
Business	-0.04006	0.024
Education	-0.04141	0.03216
Engineering	-0.47479	0.02260
Humanities	0.21827	0.02520
Science	-0.08472	0.02415
Social Sciences	0.23598	0.02356

Dummy Variables for Model				
Category	$x_{iBusiness}$	$x_{iEducation}$	...	$x_{iSocialSciences}$
Applied (Reference Category)	0	0	...	0
Business	1	0	...	0
Education	0	1	...	0
Engineering	0	0	...	0
Humanities	0	0	...	0
Science	0	0	...	0
Social Sciences	0	0	...	1

Estimated parameters			
College	Logit	Odds	Probability
Applied (Reference Category)	0	0	...
Business	-0.04006		
Education	-0.04141		
Engineering	-0.47479		
Humanities	0.21827		
Science	-0.08472		
Social Sciences	0.23598		