Categorical Data Analysis of SOTE Data

Presented to
Dr. Bee Leng Lee

Department of Statistics
San José State University

In Partial Fulfillment
Of the Requirements for the Class
Math 258

Group E
Maham Niaz, Scott Li, Yi Xiao

December 2017

Table of Contents

Introduction

The Student Opinion of Teaching Effectiveness (SOTEs) are surveys that students complete near the end of each semester in order to evaluate their instructors. Students are asked to complete one survey per course they are registered in. These evaluations are not only used by instructors to see anonymous student feedback, but also are used by staff to determine teaching performance. Thus, these ratings are taken very seriously and can influence promotion and tenure.

These ratings are controversial, since enough students may be biased or inaccurate when evaluating their instructors or the interpretation of the survey results may not be straightforward. This report analyzes data from the SOTE surveys from the 2014-2015 academic year at San Jose State University (SJSU). The goal is to determine which factors may influence these ratings and the nature of influence.

This report includes a dataset description, a brief look at missing values and response validation, the analysis of two-way contingency tables to describe the relationship between measures of student performance and overall instructor effectiveness, and lastly a logistic regression model to describe the effect of student level and college department on the probability of giving the highest effectiveness rating.

Dataset Description

The SOTE data from the 2014-2015 academic year contains data from 169429 survey forms. The surveys themselves contain 17 questions in the following order:

1. The instructor demonstrated relevance of the course content.
2. The instructor used assignments that enhanced learning.
3. The instructor summarized/emphasized important points.
4. The instructor was responsive to questions and comments from students.
5. The instructor established an atmosphere that facilitated learning.
6. The instructor was approachable for assistance.
7. The instructor was responsive to the diversity of the students in this class.
8. The instructor showed a strong interest in teaching this class.
9. The instructor used intellectually challenging teaching methods.
10. The instructor used fair grading methods.
11. The instructor helped the students analyze complex/abstract ideas.
12. The instructor provided meaningful feedback about student work.
13. Overall, this instructor’s teaching was…
14. What is your current estimate of your expected overall grade in this course?
15. You are a (freshman, sophomore, etc)...
16. Did you complete this form without undue influence from other students?
17. Did you complete this form without undue influence from the instructor?

The first 12 questions were standardized rating items on a five-point Likert scale:

Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree

Question 13 is a summary measure that is also on a five-point Likert scale, with a sixth option being N/A or no opportunity to observe :

Very Ineffective, Ineffective, Somewhat Effective, Effective, Very Effective

The data also denotes where surveys contained no response for each of these items.

Along with the survey item responses, the dataset contains 13 variables describing the survey respondents or the class.

- Official Grade
- Grading Basis
- Subject
- Level (Student Standing)
- College (Department)
- Student Level (Year in School)
- Component (Type of Course Format)
- Total Enrollment
- Enrollment Cap
- Registered Count
- Open U Count
- Instruction Mode (Instruction Format)
- Instrument (SOTE or SOLATE)

The Instrument variable does not need to be considered for analysis since it only shows the data consists of SOTE forms only and no Student Opinion of Laboratory and Activity Teaching Effectiveness (SOLATE) forms.

Main Statistical Methods, Assumptions, and Caveats

The dataset is generated in a manner consistent with a cross-sectional study because the survey is taking a snapshot of the opinions of students at a single point in time. Also, the data contains unfilled surveys, implying that the total sample size, or number of surveys, is fixed. Thus, the underlying probability structure of the data is treated as multinomial. The multinomial probability structure is versatile, since subsets of the tabulated data remain multinomial. This allows for the analysis of the association between two categorical variables. This report mostly examines the association between a variable treated as explanatory and Question 13 treated as the response.

The Chi-square test of independence is used to determine whether there is an association between the two factors. Although the assumptions of the test, especially the convergence in distribution to Chi-squared, may not always apply due to low counts for some variable combinations, the test always showed that the association is statistically significant. Thus, with this report, any analysis describes the nature of association since the presence of an association has already been determined. A multinomial probability structure also allows for the computation of several measures of association such as difference of proportions, relative risk, and odds ratios.

One caveat to the data is that it has been anonymized. Thus, individual students and the number of unique students cannot be identified. Because of this, interpretations need to be handled with care. For example, no comments can be made on proportions of students, only proportions of surveys since students may give different numbers of evaluations. However, these may approximate each other given that most students, of all levels and departments, take the same number of classes. Thus, for this report, interpretations regarding the students are drawn from interpretations regarding the surveys themselves. Another caveat is that the data is from 2014-2015 and samples a majority of students. Normally, conclusions drawn from a multinomial probability structure only apply to the time of sampling. However, interpretations from statistical inference could possibly be extended to the student population a year later or even more. But this should be done with caution. It is also questionable to extend these conclusions to student populations in other schools that use SOTEs.

Question 13

Question 13, the summary measure of teaching effectiveness, is arguably the most important question of the survey. This item is treated as the overall effectiveness of an instructor. Of the surveys that contained responses and gave a rating for Question 13, the strength of association between the first 12 questions and Question 13 is high. The Goodman Kruskal Gamma for these pairs ranged from 0.86 for Question 7 to 0.93 for Question 12. This statistic is a measure of rank correlation and measures the strength and direction of association for two ordinal variables. Thus, the pairwise associations of the questions with Question 13 are high. A Cronbach's alpha for Questions 1 to 13 is 0.97, meaning these questions have a high internal consistency. Overall, Question 13 is a good summary measure. However, this does not mean the other questions are redundant. Thus, for a complete evaluation of an instructor, all rating questions should be examined, perhaps even with a composite score if ease of evaluation is desired.

From all the surveys that gave a rating for Question 13, the median is 4 and the mean is 4.2. Thus, the average rating is quite high and is in between “effective” and “very effective”, although the mean is difficult to interpret unless it is assumed that the Likert scale has equal intervals. This may not be the case physiologically when evaluating levels of agreement. Nevertheless, there are many high ratings. Of the surveys that gave a rating for Question 13, 47.1% gave the “Very Effective” rating.

No Response Data

Only 2.7% of the surveys were empty. 76.9% of the surveys are filled out completely. 15.2% of the surveys have exactly one missing response and 7.9% have 2 or more missing responses. As expected, the most answered question is the first (1.1% no response rate). The least answered question by far is Question 7 (5.7%), possibly since students consider it vague or it requires the most thought to answer. Only 41 surveys answer the first question and skip the rest. 624 surveys skip the first twelve questions and answer the rest.

A lack of responses can hurt survey results. This data does not appear to be missing at random. If there are enough missing responses, the data could be a non-representative sample and thus conclusions made with the data could be distorted. However, an overwhelming majority of surveys (96.3%) have no more than 2 missing responses and the number of surveys is large. Thus, it can be assumed that the missing values do not harm the analysis, although a more detailed analysis of missing values should be done to confirm this. With the analyses in this report, surveys with a no response in the variables of interest are not included.

Survey Response Validation

Response validation, or checking for dishonest responses, is critical when examining the results of a survey. This is different from survey validity, which is the ability of the survey to measure what it is intended to measure. Dishonest responses are often more detrimental to survey results than missing responses. Survey respondents can be dishonest mainly by answering randomly, answering falsely, or giving the same response across all questions in order to save time and thought. Random or false responses are difficult to identify but some cases are obvious. For example, ten surveys gave the first twelve questions the lowest rating and gave Question 13 the highest rating. The occurrence of these wildly inconsistent ratings are very infrequent. The surveys with the same response level for the first thirteen questions are in Table 1 below, with their percentage of the total number number of surveys.

Table 1. Count of Surveys with Same Response Level for Questions 1 to 13

Response Level	1	2	3	4	5
Count (Percent)	642 (0.4)	256 (0.2)	5273 (3.1)	5844 (3.5)	44222 (26.1)

From this summary alone, it cannot be determined whether the survey respondents were speeding through the survey and being dishonest, although the large percentage of “all-fives” is interesting. Another opportunity to check for consistency of answers is between Question 15 (You are a (freshman, sophomore, etc)...) and the official Student Level. Out of the surveys that responded to Question 15 and have official student levels of Freshman, Sophomore, Junior, Senior, Graduate, and Credential, 21% do not match up. However, this is probably mostly due to confusion about whether these categories refer to year in school or total units taken since 96% of Freshman, 98% of Graduate students, and only 21% of Sophomores answer Question 15 consistently with their official Student Level.

Overall, dishonest responses cannot be identified with simple summary methods, except for the most extreme cases. The validity of the responses will not be mentioned with other analyses in the report.

Expected Letter Grade and Question 13

Table 2 is analyzed, which omits the Question 13 responses of No Response and N/A and the Question 14 responses of No Response and Other so that the association between expected letter grades and Question 13 can be examined. Moreover, the table is designed in such a way that it is in decreasing order for both the variables. For this analysis, the expected grade of the student is treated as an explanatory variable.

Table 2. Contingency Table for the Ordinal Data with Proportions

	A	B	C	D or F
Very Effective	43786 (0.271)	27007 (0.167)	5241 (0.032)	367 (0.002)
Effective	18583 (0.115)	23437 (0.145)	6807 (0.042)	509 (0.003)
Somewhat Effective	7055 (0.044)	11260 (0.070)	7368 (0.046)	871 (0.005)
Ineffective	1290 (0.008)	2436 (0.015)	1879 (0.012)	522 (0.003)
Very Ineffective	711 (0.004)	1225 (0.008)	889 (0.005)	397 (0.002)

A Goodman-Kruskal Gamma for this subtable is 0.43, with a 95% confidence interval of $(0.42,\ 0.43)$. This implies that there is a moderate positive association between the students opinion on effectiveness and their expected grade. In other words, students with a higher expected grade are more likely to give a higher effectiveness rating and students with a lower expected grade are more likely to give a lower effectiveness rating. The global odds ratios, shown in Table 3, can give more details on this relationship.

Table 3. Global Odds Ratios by Cutpoint Location

	A–B	B–C	C–D or F
Very Effective–Effective	2.80	3.68	5.74
Effective –Somewhat Effective	2.92	4.34	7.48
Somewhat Effective–Ineffective	3.08	4.03	9.39
Ineffective–Very Ineffective	2.85	3.80	9.67

All these values are above one, confirming the relationship evidenced by Goodman-Kruskal's Gamma and indicating that there is a consistent direction of association across the ordered categories. Interestingly, there is somewhat of a gradient from the upper left to lower right. The strongest association is between an expected grade of a D or F and the lowest effectiveness rating. The weakest association is between an expected grade of an A and the highest effectiveness rating. However, there is still a positive association here. Also, the lower a student's expected grade is, the more predictable their effectiveness rating is.

Thus, a positive association exists between a student's expected grade in the course, and their rating of the instructor's teaching. This association gets stronger with a lower expected grade. The lower a student's expected grade is, the more likely they are to give a lower rating.

Difference Between Official and Expected Grades and Question 13

In the data set, the Official Grade has 20 levels: A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F, Credit, No Credit, Incomplete, Withdrawal, Withdrawal Unauthorized, Report Delayed, and Report in Progress. The expected grades (Question 14) has 5 levels: A, B, C, D and F, and No

Response. In order to compute a difference measure between the official and expected grade, and apply one-to-one mapping with the response, only the surveys with letter grades and a response to Question 14 were examined. Also, the Official Grade needed to be collapsed into the 4 letter-grade levels of Question 14. Then, with A being assigned 1, B = 2, C = 3, and D and F = 4, the difference (official minus expected grade) was computed. For the analysis of the association between grade difference and Question 13, only surveys that responded to Question 13 are examined. The total number of observations is 156813. Table 4 is the contingency table with grade difference as the rows and Question 13 responses as the columns.

The difference of the students rating is looked at as pessimistic and optimistic. Pessimistic refers to a student who's expected grade is lower than the Official Grade and optimistic student refers to a student who's expected grade is higher than the Official Grade. Negative differences denote the pessimistic students and positive denote the optimistic students. A difference of zero means that the expected grade was equal to the Official Grade and in a sense, those survey respondents predicted their grades correctly or have realistic expectations. Most students fall in this category. There are more optimistic students than pessimistic students.

Table 4. Grade Difference and Question 13

Actual-Expected Grade	**Very Ineffective**	**Very Ineffective**	**Somewhat Effective**	**Effective**	**Very Effective**
-3	11	12	2	5	5
-2	90	145	509	260	252
-1	666	1394	5507	8306	8525
0	1695	3306	14662	30070	48467
1	610	1035	4678	8504	14402
2	64	99	422	772	1956
3	1	3	27	28	323

The Goodman-Kruskal Gamma is computed to be 0.15. This number implies that there is a small positive association. This association is better described with specific measures. Since both variables are ordinal, global odds ratios, shown below, can be used to describe the nature of association. The table separates pessimistic and optimistic students for clear interpretation since the students who predict their grade correctly can be used as the point of comparison. Table 5 shows the global odds ratios for the split groups, with the cut-points for grade difference.

Table 5. Global Odds Ratios for Pessimistic and Optimistic Grouping

Pessimistic					Optimistic				
-3/-2	22.70	30.62	8.57	5.16	0/1	0.84	0.92	0.94	1.05
-2/-1	4.32	4.09	5.16	3.50	1/2	1.03	1.16	1.28	1.65
-1/0	1.75	1.85	1.92	1.88	2/3	7.03	5.19	2.89	5.57

The global odds ratios for pessimistic students are consistently greater than 1 here, which means the lesser the difference between actual and expected grade, the higher the teacher's ratings.

This result is interesting because instructors who are strict graders in exams and homeworks to enhance learning, but pass the students in the end with better grades may be given lower ratings. For the optimistic students, the global odds ratios are close to 1 but increase after the difference of 1. This means that as the difference between actual and expected grades increases, the less likely are the students to give teachers lower ratings. Teachers can possibly make the students believe that they will get good grades but give them low grades in the end, and can increase their chances of good ratings and also be seen as very strict.

To further describe the nature of association between the grade difference and how students rate the teachers, relative risk can be used after collapsing the data into 2x2 tables shown in Table 6 and Table 7.

Table 6. Pessimistic Surveys

Grade Difference	Somewhat Effective or Lower	Effective or Very Effective
0	19663	78537
-1,-2,-3	8336	17353

The estimated relative risk is 0.618. The probability of giving a lower rating to the instructor when the difference between expected and observed grades is 0 is 0.618 times the probability when the difference is negative. Pessimistic students are about 1.62 times (62%) more likely to give teachers lower ratings.

Table 7. Optimistic Surveys

Grade Difference	Somewhat Effective or Lower	Effective or Very Effective
0	19663	78537
1,2,3	6939	25985

The estimated relative risk is 0.950. The probability of giving a teacher lower ratings when the difference in expected and official grades is 0 is 0.95 times the probability of giving a teacher lower ratings with optimistic students. This difference is small.

The Bonferroni corrected confidence intervals at a 95% level are:

Pessimistic: (0.602, 0.633). Inverting this interval (1.58, 1.66) means the pessimistic students are estimated to be anywhere from 58% to 66% more likely to give bad ratings to a teacher compared to neutral students.

Optimistic: (0.924, 0.977). Inverting this interval (1.02, 1.08) means the optimistic students are estimated to be 0.08% to 0.02% more likely than the neutral students to give a bad rating to the teacher. This difference is negligible.

With these 2x2 tables, the pessimistic students tend to rate the teachers lower. However, the difference for optimistic students is not that much.

Binary Logistic Regression Model

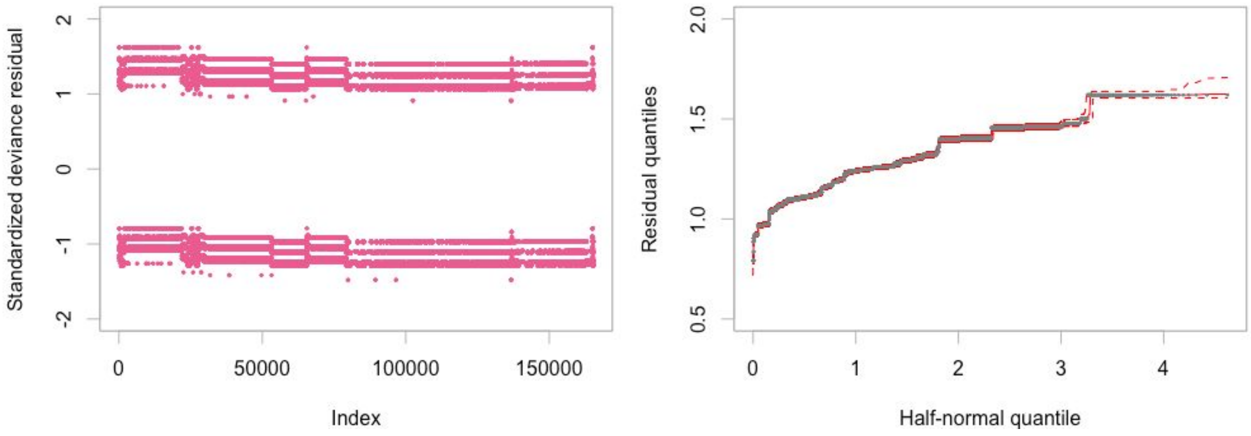
The logistic regression model not only helps to examine the extent by which certain factors can influence instructor ratings, but can also be used to predict the probability of getting a high rating.

Question 13 is turned into a binary response variable by taking all surveys that gave a response and treating a rating of “Very Effective” as “1” and the other ratings as “0”. This is done in order to distinguish the highest rating, which instructors need to stand out and achieve a high average score, from the other ratings. This division is logical, given that there are a large number of “Very Effective” ratings in the first place.

A hypothesis of interest is that student level and college level plays an important role in student satisfaction or students ratings of the instructors. Freshmen students may tend to be more lenient, while senior or graduate students are more strict, perhaps because of experience. Another categorical variable of interest is the College. It has been hypothesized that students in the non-STEM related colleges give better ratings to the teacher, while the students enrolled in STEM related colleges are more strict in their evaluations due to the stress level of different subjects.

Originally, continuous variables measuring class size or proportion of the class filled were considered but exploratory plots showed that these variables gave poor separation of the binary response with a Logit link. Thus, only the two categorical predictors are included without interaction terms. This model assumes that the variables are uncorrelated and that there is linearity between the predictors and the log odds. The adjusted symmetric uncertainty coefficient between Student Level and College is about 0.10 or 0.08 (subsetting out the zero counts or replacing the zero counts with a value of one was required to compute this number). This is not a strong relationship so multicollinearity will not be an issue. The second assumption is more difficult to address although several model diagnostics are examined.

Figure 1. Binary Logistic Regression Diagnostic Plots



Two common diagnostic plots for binary logistic regression are shown in Figure 1. The standardized deviance residual plot (on the left) shows distinct bands since there are only categorical predictors in the model. There are several dozen outliers, judging from the points that appear to stand alone, but that is expected with 165413 observations. Most of these outliers are from the "Undeclared" department. Surveys from this department have the largest proportion of "Very Effective" (58.6%) ratings so they could be more difficult to classify. The Half-normal quantile plot (on the right) shows no indication of a lack of fit, since most points follow the median and all lie within the simulated envelope. The Hosmer-Lemeshow C and H statistics have p-values of 0.016 and 0.18 respectively. At a 0.05 level, there could be a concern with the goodness-of-fit of the model since the C statistic is significant. Overall, there is some evidence that the model is not a good fit based on one test, but with the plots and tests considered, the fit is probably adequate. The following equation calculates the probability of giving a very effective rating to an instructor based on college and student level.

$$\widehat{\pi}(x) = exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_{13} x_{13}) / (1 + exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_{13} x_{13}))$$

This model has 14 parameters. There are 8*7 = 56 total combinations: 8 colleges and 7 student levels. The parameter estimates are presented in Table 8.

Table 8. Binary Logistic Regression Parameter Estimates

	Parameter	β Estimate	Standard Error
	Intercept	-0.33	0.019
Student Level (Freshmen as the Baseline)	Sophomore	0.27	0.033
	Junior	0.35	0.018
	Senior	0.49	0.017
	Postbaccalaureate	0.13	0.17
	Credential	0.44	0.046
	Graduate	0.48	0.021
College (Applied Sciences and Arts as the Baseline)	Business	-0.35	0.018
	Education	0.0042	0.024
	Engineering	-0.66	0.018
	Humanities	0.10	0.018
	Science	-0.30	0.018
	Social Sciences	0.030	0.017
	Undeclared	0.51	0.19

The Postbaccalaureate, Education, and Social Sciences parameters are statistically insignificant (no different from zero according to a Wald test). Thus, care should be taken when predicting probabilities with these categories. When these parameters are close to zero in this model, it means they give similar probabilities to their respective baseline categories.

Instead of seeing the effect of each parameter on the odds when compared to the baseline, a clear way to describe the model is to tabulate the estimated probabilities of giving a rating of "Very Effective" for some combinations of interest (Table 9).

Table 9. Estimated Probabilities of "Very Effective" Rating

Student Level	College Departments				
	Engineering	Business	Science	Applied	Humanities
Freshman	0.27	0.33	0.35	0.42	0.44
Sophomore	0.32	0.40	0.41	0.48	0.51
Junior	0.34	0.42	0.43	0.50	0.53
Senior	0.38	0.45	0.46	0.54	0.57
Graduate	0.37	0.45	0.46	0.54	0.56

It is clear that within these five departments, the estimated probability of giving a rating of "Very Effective" increases with an increase in the student level. This could be due to many factors. Perhaps instructor quality increases when teaching more specific classes instead of classes for general education requirements, since the jump in probability from Freshman to Sophomore seems to be the greatest (a 5 to 7 percentage point increase). Maybe an increase in student experience means they tend to evaluate instructors by a different standard.

Judging by this table and the Engineering parameter (-0.66), the probability of a student rating the instructor as "Very Effective" in the Engineering department is the lowest out of all the departments. The probability of giving a "Very Effective" rating for graduate students in the Humanities is 0.19 points more than the probability for graduate students in Engineering. There appears to be a difference between departments, so it is important to evaluate instructors based on department averages as well as overall averages.

Conclusion

The SOTEs are an important means of judging instructor effectiveness and often influence tenure and promotion. Therefore, this paper provides some analysis of SOTE data to understand the nature of association of variables related to student's characteristics with the instructor's effectiveness. For analysis, Question 13 was used as a response variable since it is a good summary of the instructor's positive characteristics. The ratings tend to increase with an increase with a student's expected grade. The effect of the difference in expected and official grades on the student's rating was also examined. The students whose expected grades are lower than official grades tend to rate the teachers lower compared to students who predict their grade accurately. The effects of student level and college were also explored with binary logistic regression. The original hypothesis, which said that as the student level increases, the instructor's ratings get better, did not hold after running logistic regression. The students who are older or more experienced are more likely to give the highest rating. Also, there is a difference among STEM and non-STEM colleges as suspected. Students enrolled in STEM colleges are more strict in rating the professors. Overall, there are several student factors that can influence instructor ratings at SJSU. Thus, there is more to a rating than the overall quality of an instructor's teaching and these covariates need to be taken into account when evaluating each instructor. Further investigation should be done on how students interpret the survey questions, such as Question 14, the extent of the bias these factors give, and whether they can be adjusted for.