Evidence on relationships between teaching evaluations and ratemyprofessors teaching tags

Terrance Jalbert
University of Hawaii Hilo

ABSTRACT

This research examines relationships between teaching tags and teaching evaluations. The paper constitutes one in a series of papers by the same author that examines Ratemyprofessors.com data. Since 2014, Ratemyprofessors' reviewers can assign tags that provide detailed comments about the class and professor. The analysis here examines how these student-assigned teaching tags relate to teaching evaluations and the willingness of a student to take another course from the professor. This paper analyzes data for 202 business professors. Results identify variables important and unimportant in determining teaching ratings. Professors might focus efforts on critical areas identified here to improve their teaching ratings.

Keywords: Determinants of Teaching Evaluations, Ratemyprofessors, Teaching Tags

INTRODUCTION

Student evaluations of teaching provide a useful method of assessing professor performance. These evaluations allow for standardized comparison across professors and across universities. Evidence indicates increasing popularity of teaching evaluations for determining instructor performance. Miller and Seldin (2014) conducted a survey of 401 accreted four-year liberal arts colleges. Results indicate that 99.3 percent of these colleges consider classroom teaching a major factor in evaluating faculty performance. Moreover, 94.2 percent of colleges use systematic student evaluations in determining teaching performance. Student evaluations exceed the popularity of second place variable, chair evaluation, by more than fifteen percent. Emphasis on student evaluations increased by more than six percent from 2000 to 2010.

Given the importance of teaching evaluations, it is valuable to study determinants and potential biases in these evaluations. One difficulty associated with conducting studies on teaching evaluations occurs because the evaluations are generally not publicly available. Thus, most studies examining evaluations include data limited to a single university. A recently developed instructor rating tool, called Ratemyprofessors (RMP) allows students to make public reviews of faculty. As of June 27, 2019, the service reports ratings for 1.7 million professors from over 7,500 schools, including more than 19 million individual ratings (Ratemyprofessors.com, 2019).

The Ratemyprofessors tool enjoys high levels of popularity and respect among students. Brown, Baillie and Fraser (2009) provide an excellent student survey regarding Ratemyprofessors. Results show that 83 percent of students have visited RMP and 36 percent have rated a professor on RMP. Students indicated that RMP ratings constitute an honest and representative measure of teaching abilities. Indeed, 96 percent of participants believed that students provide more or equally honest reviews in their RMP ratings than standard teaching evaluations. Moreover, 81 percent of students believed RMP ratings better or equally represent instructors' performance than standard teaching evaluations. Students indicated substantial intentions to use RMP ratings when making academic decisions. Some 71 percent of students avoided taking an instructor because of their RMP rating. Results show significant correlation between RMP and standard teaching ratings. However, there remains some unexplained differences between the two ratings methods. Their results show that standard teaching evaluations tend to be higher than RMP ratings.

Since September of 2014, RMP reviewers can assign tags to teaching evaluations. Students can select up to three tags out of a list of 20 provided by Ratemyprofessors (Ratemyprofessors, 2019). The Tags function allows students to assign indicators to an evaluation that provide insights into important determinants of their evaluation. This study uses RMP data with a focus on the relationship between Tags and overall teaching evaluations. By examining these data for multiple professors and multiple schools, this study provides a national study of teaching evaluation determinants.

This paper represents the third in a series that relies on variations of the same dataset. In an earlier paper, Jalbert (2019a) combines Ratemyprofessors and Social Science Research Network Data (SSRN) for 300 business professors to examine relationships between professor teaching and research ratings. Results indicate no relationship between teaching and research ratings. The author argues that teaching and research duties represent separate and distinct activities. Results further indicate that course difficulty, expertise area, experience, tenured status and Hotness significantly impact teaching ratings. In addition, rating profiles differ significantly between public and private schools. Jalbert (2019b), examined the extent to which professors who excel at both teaching and research exits. Results indicate that few professors exist who achieve excellence both in teaching and research. Fewer than 25 percent of professors placed in the top 50 percentile on both measures. Moreover, many seriously deficient professors exist in the areas of teaching, research or both. Results show 50 percent of all professors falling in the lower 30th percentile for teaching, research, or both.

The remainder of the paper transpires as follows. In the next section we provide a review of the extant literature. The data and methodology section presents data utilized in the study, provides some summary statistics and a description of how data examination proceeded. Next, the approach provides the empirical results and discussion. The paper closes with some concluding comments and some recommendations for professors.

LITERATURE REVIEW

A large body of research examines teaching evaluations. Much of this research examines how various demographic factors relate to teaching evaluations. Marks (2000), used structural equation modeling to examine student evaluations. He identified five latent variables in student evaluations as follows: organization; workload and difficulty; expected grades and fairness of grading; instructor liking and concern; and perceived learning.

Mixed evidence exists on differences in teaching ratings by gender. Some authors find that female professors receive lower teaching evaluations than male professors. Wagner, Rieger and Voorvelt (2016) found that women are eleven percent less likely than men to earn teaching evaluations that meet the minimum requirements for tenure and promotion. Miles and House (2015) found that female and males are generally competitive, but that females earn noticeably lower scores in large classes. Mengel, Sauermann and Zölitz (2018) examined a sample including nearly 20,000 teaching evaluations. They found that women received systematically lower teaching evaluations than their male counterparts. Bosow (1995) evaluated student evaluations by gender of both the student and professor. She found no impact of student gender on male professor's evaluations. However, male students ranked female professors lower than female students. Similarly, Centra and Gaubatz (2000) found that female students consistently give higher evaluations to female professors. Jalbert (2019a and 2019b), found that female professors receive significantly higher teaching evaluations than male professors. He cautions readers on interpreting results from these studies. He notes that his results, and those from other similar studies, could be driven by either rater bias or gender-based differences in teaching capabilities. In a rare paper to address both issues, Boring, Ottoboni and Stark (2016) find that females receive lower teaching evaluations and further find that these differences are not reflective of differences in teaching quality.

A small body of research examines the extent to which class size impacts teaching evaluations. McPherson (2006) examined principles level courses finding that class size negatively affects teaching evaluations. Feldman (1984) found a small inverse relationship between class size and students' overall evaluation of the teacher and course. Miles and House (2015) found that class size impacts teaching evaluations. Linsky and Strauss (1975) found a negative relationship between overall teaching ratings and university enrollment. This result suggests that professors from small schools earn higher teaching evaluations than those from large schools. However, Jalbert (2019a) found no relationship between university size and teaching evaluations.

A stylized fact in higher education suggests that course grades significantly impact teaching evaluations. McPherson (2006) found that expected grades significantly impact student teaching evaluation scores. Similarly, Bilgen, Susanh and Kaytaz (2015) found positive associations between student teaching evaluations and grades in Turkey. Miles and House (2015) examined more than 30,000 evaluations for 255 professors. Their results show that higher teaching evaluations associate with higher expected grades. Isley and Singh (2005) examined expected grades and cumulative grades as they relate to teaching evaluations. They found the gap between expected grade and cumulative grades significantly explains teaching evaluations with a negative sign. This implies that larger discrepancies lead to lower evaluations. The authors argue this gap constitutes a more appropriate variable than expected grade in explaining teaching evaluation scores. Brown, Baillie and Fraser (2009) found that difficulty ratings on RMP evaluations relate to teaching evaluations in a stronger way than for standard teaching evaluations.

A small body of literature examines the role of humor in education and in teaching evaluations. Garner (2006) examined evaluation scores for 117 undergraduate students in a distance education format. He found that courses with humor received significantly higher evaluations than those without humor. Moreover, students retained more information when humor was incorporated into the course. Banas, Dunbar, Rodriguez and Liu (2010) provided a review of the existing literature and summarize the results by noting the use of nonaggressive appropriate humor in the classroom associates with a more relaxed and interesting learning environment, higher instructor evaluations, stronger motivations to learn and more course enjoyment. Sojka, Gupta and Deeter-schmelz (2010) found that faculty believe easy and entertaining instructors receive higher teaching evaluations.

An interesting study by Leis and McKinzie (2019) examined the impact of industry experience on teaching evaluations. They examine 355 sets of teaching evaluations and identified a positive relationship between industry experience and instructor evaluations. Interestingly, they find negative relationships between years of teaching experience and three evaluation criteria: defining expectations and objectives, effective communication and availability of professor.

DATA AND METHODOLOGY

This paper utilizes data from Ratemyprofessors.com. Data collection began with the sample used in Jalbert (2019a) and Jalbert (2019b). The initial sample included 300 business professors from 104 universities, who were both listed in the Social Science Research Network (SSRN) website and had 10 student reviews on Ratemyprofessors. Collected data includes the teaching rating and difficulty rating of each professor from RMP. Data collection also involved recording the number of student reviews the professor realized and the percentage of students who indicated a willingness to take another class from the professor since May 25, 2016. The analysis also involved collecting Tag data. For evaluations completed after September of 2014, students may add up to three Tags to their review indicating how the students describe the professor and course. Ratemyprofessors aggregates these Tags and provides totals for professors. The following list reports the twenty candidate Tags students may select from: Gives Good Feedback, Respected, Lots of Homework, Assessible Outside Class, Get Ready to Read, Participation Matters, Skip Class? You Won’t Pass, Inspirational, Graded by a Few Things, Test Heavy, Group Projects, Clear Grading Criteria, Hilarious, Beware of Pop Quizzes, Amazing Lectures, Lecture Heavy, Caring, Extra Credit, So Many Papers and Tough Grader (Ratemyprofessors.com, 2019).

To supplement professor data, the analysis involved gathering data on the university where the professors teach. Collected data includes *University Quality*, *University Reputation* and *University Average Ranking* (Each with a scale from 0-5 with 5 equaling the highest score). Finally, we identified data on the *Number of Professors* reviewed at the university and use this as a measure of university size. The *Public vs Private* variable indicates 1 if the school is public and 0 if private. The *Professor Gender* variable indicates 1 for female and 0 for male.

Next, the process involved reducing the dataset on two metrics to assure that the dataset included only recently active professors. To eliminate obsolete observations, data reduction involved eliminating professors whose most recent Ratemyprofessors.com rating occurred prior to January 1, 2017. This process reduced the dataset by 94 observations to 206 observations. Next, the method removed four observations that included no reported Tags. This resulted in a final dataset of 202 professor observations for analysis including 7,563 total evaluations implying an average of just over 37 evaluations per professor.

The analysis considered five measures of teaching evaluation. *Raw Evaluation* as reported by RMP constitutes the first measure. The process continued by making three enhancements to this measure. Given the stylized fact that course difficulty impacts teaching evaluations, the second measure of performance combines teaching ratings and course difficulty measures. This paper equally weights the two measures to arrive at a new measure, *Weighted Evaluation*, in a manner analogous to Jalbert (2019a and 2019b). Equation 1 shows the calculations of this measure.

Next, the approach considered the possibility that students from different universities evaluate professors in different ways. The analysis here uses RMP *University Average Rating* to standardize the Raw evaluations across universities. Equation 2 shows the *Standardized Evaluation* technique. By multiplying the relative evaluations by 5 in Equation 2, the *Standardized Evaluations* utilize the same scale as raw and weighted evaluations. The fourth approach both weighted and standardized the data, *Standardized and Weighted Evaluation*, as shown in Equation 3. These first four measures have values ranging from 1-5 with 5 equaling the highest. A final teaching evaluation measure uses a new measure of teaching performance available from RMP whereby students indicate if they would take the professor again. The *Willingness to Retake* variable equals the percentage of students who indicate they would take the class again.

$$Weighted\ Evaluation = \frac{Raw\ Evaluation + Course\ Difficulty}{2} \tag{1}$$

$$Standardized\ Evaluation = \frac{Raw\ Evaluation}{University\ Average\ Evaluation} X\ 5 \tag{2}$$

$$Standardized\ and\ Weighted\ Evaluation = \frac{Standardized\ Evaluation + Course\ Difficulty}{2} \tag{3}$$

The data included observations from 91 universities with ten universities producing four professors, 25 universities producing 3 professors, 31 universities producing 2 observations and 25 universities producing 1 professor. The *Number of Professors* evaluated at the sample universities ranged from under 350 to more than 8,500 indicating considerable size diversity among the sample universities. Data includes 57 finance professors, 52 accounting professors, 52 economics professors, 28 management professors and 13 professors classified as others.

As noted earlier, students may select up to three tags to accompany their evaluation. Table 1 (Appendix) shows available tags from which student can select. Aggregate Responses indicates the total number of responses received by all sample professors. Recall that the total number of evaluations in the dataset equaled 7,563. The largest number of responses equals 745 for Tough Grader followed by Skip Class You Won't Pass and Respected with 596 and 461 responses respectively. The next two columns show the number of professors having at least one response and the number of professors without a response for each Tag. The most common Tag was Tough Grader. The least common tag was So Many Papers. Percent with Responses indicates the percentage of professors with one or more responses for a Tag with values ranging from 6.93 percent to 77.72 percent. The column Mean Number of Responses shows the average number of responses for those professors who received the Tag. The largest mean number of responses equals 4.745 and the minimum equals 1.071.

RESULTS

The analysis begins with single regressions of each of the twenty-eight independent variables individually on the dependent variables. The approach used ordinary least squares regression and estimated the following three models for each independent variable where ε is a random error term:

$$Raw\ Evaluation = \alpha + \beta_1(Independent\ Variable) + \varepsilon \tag{4}$$

$$Weighted\ Evaluation = \alpha + \beta_1(Independent\ Variable) + \varepsilon \tag{5}$$

$$Willingness\ to\ Retake = \alpha + \beta_1(Independent\ Variable) + \varepsilon \tag{6}$$

Table 2 (Appendix) shows the results. Panel A provides results for the Raw Evaluation dependent variable. Panel B and C show results for the Weighted Evaluation dependent variable and Willingness to Retake dependent variable respectively. Raw Evaluation results reveal eight significant independent variables at the one percent level: Amazing Lectures, Hilarious, Respected, Caring, Inspirational, Assessable Outside Class, Course Difficulty and Good Feedback. An additional four variables show significance at the 5 percent level: Get Ready to Read, Clear Grading Criteria, Expertise Area and Public Versus Private.

In each case, the data revealed coefficients consistent with the expected sign as noted in the column labeled Expected Sign (E.S.) Each variable produced a positive response with the exceptions of Get Ready to Read, Course Difficulty and Public versus Private. The data here confirms the common conception that course difficulty affects teaching evaluations. R2's on the regressions range from 0.0004 to 0.1707 with course difficulty and amazing lectures producing the highest R2's.

Regressions on Weighted Evaluations produced similar results. Again, twelve independent variables produce significant results including: Amazing Lectures, Hilarious, Respected, Caring, Inspirational, Assessible Outside Class, Skip Class You Won't Pass, Good Feedback, Expertise Area, School Reputation, Public versus Private and Number of Professors. R2 values ranged from 0.0000 to 0.642 with Amazing Lectures, Respected, and Good Feedback producing the highest R2. Two significant variables in the Raw Evaluations did not reach significance in the Weighted Evaluation results: Get Ready to Read and Clear Grading Criteria. In addition, Course Difficulty was not incorporated as an explanatory variable in the Weighted Evaluation examinations. Three insignificant variables in the Raw Evaluation results achieved significance in the Weighted Evaluation results: Skip Class you Won’t Pass, School Reputation and Public Versus Private.

Results reveal robustness to changes in dependent variable specifications. Standardized Evaluations (Equation 2) and Weighted Standardized Evaluations (Equation 3) produced similar results to the raw and weighted variables presented here. For this reason, the tables here exclude these results.

The analysis turns to an examination of the dependent variable Willingness to Retake. Recall the Willingness to Retake variable indicates if the student indicate they would retake the class. Many observations did not include data for this variable. For this reason, regressions on Retake include 108 observations. Results show the variables Amazing Lectures, Hilarious, Respected, Caring, Assessible Outside Class, Get Ready to Read, Course Difficulty, Text Heavy, Clear Grading Criteria, Pop Quizzes and Professor Gender significantly explain the variable Would Retake. The positive coefficient on Professor Gender reveals that students indicate a higher desire to retake female professors.

The approach continues with multiple regression analysis that incorporates all twenty-eight independent variables. We estimate the following equations using ordinary least squares regression for Raw Evaluations:

$$\begin{aligned}&Raw\ Evaluation = \alpha + \beta_1(Amazing\ Lectures) + \beta_2(Hilarious) + \beta_3(Respected) + \\ &\beta_4(Caring) + \beta_5(Inspirational) + \beta_6(Assessible\ Outside\ Class) + \beta_7(Lecture\ Heavy) + \\ &\beta_8(Participation\ Matters) + \beta_9(Skip\ Class\ You\ Won't\ Pass) + \beta_{10}(Group\ Projects) + \\ &\beta_{11}(Get\ Ready\ to\ Read) + \beta_{12}(So\ Many\ Papers) + \beta_{13}(Lots\ of\ Homework) + \\ &\beta_{14}(Course\ Difficulty) + \beta_{15}(Test\ Heavy) + \beta_{16}(Clear\ Grading\ Criteria) + \\ &\beta_{17}(Graded\ by\ a\ Few\ Things) + \beta_{18}(Tough\ Grader) + \beta_{19}(Good\ Feedback) + \\ &\beta_{20}(Pop\ Quizzes) + \beta_{21}(Extra\ Credit) + \beta_{22}(Expertise\ Area) + \beta_{23}(Gender) + \\ &\beta_{24}(University\ Quality) + \beta_{25}(University\ Reputation) + \\ &\beta_{26}(University\ Average\ Rating) + \beta_{27}(Public\ vs\ Private) + \\ &\beta_{28}(Number\ of\ Professors) + \varepsilon\end{aligned} \tag{7}$$

The Weighted Evaluation regressions, while otherwise identical, exclude the Course Difficulty independent variable. The regression on Willingness to Retake utilizes the same independent variables noted in Equation 7.

Table 3 (Appendix) shows the results. Results for Raw Evaluations, presented in Panel A, show twelve significant variables. Interestingly, Hilarious, Caring and Assessible Outside Class do not reach the 10 percent significant level in the multiple regression. However, some insignificant variables in the single regressions show significance in the multiple regression. These variables include: Lecture Heavy with a negative coefficient, Graded by a Few Things with a negative coefficient, Professor Gender with a negative coefficient and School Quality with a negative coefficient. The resulting R2 equals 0.5196 and the Adjusted R2 equals 0.4419.

The Weighted Evaluation examination, Panel B, produce similar results however, the variables Clear Grading Criteria, School Quality and School Reputation do not reach a 10 percent significance level. The R2 and Adjusted R2 equal 0.3438 and 0.2420 respectively.

Willingness to Retake (Panel C) produces interesting results that differ substantially from the previous results. Amazing Lectures, Hilarious, Caring and Inspirational do not achieve significance in the model. Willingness to Retake reveals a positive association with Respected, Assessible Outside Class, Clear Grading Criteria, Extra Credit, and School Average Rating. Willingness to Retake reveals a negative relationship with Lecture Heavy, Get Ready to Read, Test Heavy, and Tough Grader. The regression results in an R2 of 0.6071 and an adjusted R2 of 0.4679.

Next, the approach applied stepwise regression to the model. Stepwise regression objectively selects explanatory variables for inclusion in the model that increase the model fit. Several variable selection methods exist. The analysis here uses forward selection which involves adding variables to the model in each stage that meet the inclusion criteria. The process stops when no further variables meet the criteria. The criteria require a ten percent significance level for inclusion into the model.

Table 4 (Appendix) shows the stepwise regression results. Panel A shows results for Raw Evaluations. Results indicate that eleven variables met the ten percent significance criteria for inclusion in the model. Six variables show significance at the one percent level, three at the five percent level and two at the one percent level. As one might expect, course difficulty produced the highest Partial R2 at 0.1707. Amazing Lectures provides a partial R2 of 0.0880. Lecture Heavy, Caring and Public vs Private and Graded by a Few Things round out the one percent significance variables. The full model R2 reaches 0.4622.

Panel B shows results for Weighted Evaluations. As noted earlier, the approach excludes Course Difficulty from these regressions. Amazing Lectures, Graded by a Few Things, Assessible Outside Class and Public vs Private achieve significance at the one percent level. School Average Rating enters the model indicating some significant differences in evaluations across schools. The full model produces a lower Model R2 of 0.3126, a value lower than the Raw Evaluation results. Interestingly, Caring was an important contributor in the Raw Evaluation results but did not enter the Weighted Evaluation results. Similarly, Gender entered the Weighted Evaluation results, but was not significant in the Raw Evaluation results.

Panel C shows the Willingness to Retake results. The results here differ substantially from the Raw and Weighted Evaluation results. Test Heavy, which did not enter the Panel A or B results, entered the model here in Stage 2. It produced a partial R2 of 0.1370 indicating the type of testing impacts Willingness to Retake. Another interesting finding shows that Amazing Lectures and Course Difficulty did not enter the Willingness to Retake model. The full model resulted in a Model R2 of 0.5502.

Next we aggregate the results to glean collective inferences. Examining the combined results in this fashion allows for identification of independent variables that exhibit robustness to measurement differences. Table 5 (Appendix) shows the combined results. Panel’s A and B report results for the Single and Multiple regressions respectively. A positive notation indicates the variable entered the model significantly with a positive coefficient. A negative notation indicates the variable entered the model significantly with a negative coefficient. Variables without a notation did not enter the model. Panel C reports results for the stepwise regression. The figure in the cells show the sequence in which the variable entered the model.

Not surprisingly, Amazing Lectures shows positive significance in each simple and two multiple regressions. It also entered two stepwise regressions. The only regressions that Amazing Lectures did not enter as significant was the multiple and stepwise regressions on Willingness to Retake. This suggests the obvious path of improving lecture quality for teaching evaluation improvement. Hilarious generated significant positive results in the simple regressions for each dependent variable. However, it did not result in significant coefficients in the multiple or stepwise regressions. This finding suggests that research on the relationships between humor and teaching evaluations should carefully consider appropriate control variables.

Respected constituted the only significant variable in every regression and entered with a positive coefficient in each case. Clearly, Respected impacts teaching evaluations. Nevertheless, it is difficult to identify specific actions a professor might accomplish to increase the amount of respect they enjoy. This finding suggests the need for additional research to identify variables related to students' respect of a professor.

Caring yielded significant results in the single regressions with a positive sign and in two of the stepwise regressions. However, results did not reach ten percent significance in the multiple regressions. Overall, it appears advantageous for professors to increase their level of caring. Inspirational was significant in two of three single and multiple regressions, with insignificant results in both cases for the Retake variable. Results produced mixed signs on the Inspirational variable with positive coefficients in the single regressions, but negative coefficients in the multiple regressions. This variable entered each of the stepwise regressions late in the sequence and with marginal significance. While the result is interesting, it remains difficult to pinpoint specific actions a professor might take to increase perceived inspiration levels.

Accessible Outside of Class delivered significantly positive coefficients in each single regression and in one multiple regression. The variable entered each stepwise regression at stages up to third place. Students clearly value accessibility. Moreover, accessibility falls directly under control of the professor. For this reason, improving accessibility appears beneficial for professors seeking higher evaluation scores.

Lecture Heavy generated significant negative coefficients in each multiple regression but did not achieve significance in the single regressions. It entered each stepwise regression in the earlier stages. It appears that including other classroom activities, in addition to lectures, might produce higher teaching evaluations. Results show little effect of Participation Matters, Skip Class you Won't Pass and Group Projects on teaching evaluations.

Get Ready to Read yielded significant negative coefficients in two single regressions and one multiple regression. It also entered two stepwise regressions. The result implies that professors should carefully select reading materials for their courses and might eliminate noncritical reading. The variables So Many Papers and Lots of Homework did not produce any significant results.

Course Difficulty produced significant coefficients for both independent variables, but only in the Raw Evaluation multiple regression. It was the most significant item in the Raw Evaluation stepwise regression but did not enter the other stepwise regressions. The negative coefficient indicates a clear path for professors to increase their evaluation scores.

Next, consider grading techniques used by the professor. Results show Test Heavy impacts Willingness to Retake negatively but did not show an impact for the other dependent variables. Clear Grading Criteria, produced positive coefficients for Raw Evaluations and Willingness to Retake regressions, but did not show significance in any Weighted Evaluation regressions. Graded by a Few things resulted in negative coefficients in the raw and weighted multiple regressions and entered each of the stepwise regressions. Tough Grader generated a negative coefficient for Willingness to Retake and entered two stepwise regressions. Good Feedback yielded positive coefficients for the Raw and Weighted single and multiple regressions and entered the stepwise regressions for these two dependent variables. Pop Quizzes showed a negative coefficient on the Willingness to Retake single regression but did not result in a significant coefficient for any other model. Extra Credit returned only one negative significant coefficient. The combined evidence suggests that professors wishing to improve their evaluations should focus heavily on specifying and adhering to unambiguous grading criteria. They should work to make their examinations representative of all materials covered. Finally, providing good feedback leads to better teaching evaluations.

Professor Expertise Area yielded significant results with mixed coefficient signs. These results suggest that further research examining evaluations by area, using a larger dataset, might produce interesting results. Similarly, Professor Gender produces coefficients with mixed signs, again suggesting additional research to further understand the relationship. University Reputation, Average University Rating and Number of Professors produced sporadic significant results. The variability in those results makes it difficult to make inferences. Negative coefficients on the Public vs Private single regressions and the variables entry into two stepwise regressions indicates systematic differences between public and private university evaluations. Public university professors appear to earn lower evaluations than private school professors. It remains unclear if these differences imply private schools provide better professors to their students or there exists systematic bias in evaluations. Future research might shed additional light on this issue.

CONCLUDING COMMENTS

This paper examines determinants of business professor teaching evaluations. The examination utilizes newly available data from Ratemyprofessors.com. Ratemyprofessors.com provides a public forum for evaluating professors that is standardized across professors, universities and nations. A recently added feature of Ratemyprofessors.com allows students to incorporate Tags into their evaluations. These Tags allow students to identify selected characteristics of the professor and course that influenced their evaluation. This paper examines relationships between these Tags and associated teaching evaluations. The paper utilizes data for 202 business professors.

Results indicate some variables impact teaching evaluations, while others produce limited or no evidence of an impact on evaluations. The analysis here confirms that course difficulty impacts teaching ratings in a negative way. As one would expect higher quality lectures relate to higher teaching evaluations. While respected professors receive significantly higher evaluations, it remains unclear any specific actions professors might undertake to improve in this category. Evidence indicates that professors who make themselves more available outside of class can improve their teaching evaluations. Providing clear grading criteria and good feedback might further lead to higher teaching evaluations. Professors should manage required reading to minimize student workload whenever possible. Professors wishing to improve their teaching evaluations should focus on these areas as well as other significant areas noted in this research.

Like all research, this paper includes some limitations. These limitations include the relatively small dataset utilized in the paper. Future research might verify the results here when additional data becomes available through RMP. Further, a considerable body of literature suggests that female and male students evaluate professors differently. RMP does not report evaluator gender, making it impossible to assess these effects in the current research. Students may only list three tags for a professor and students do not rank the Tags by importance. Enhanced data on these metrics might provide new and interesting results.

REFERENCES

Banas, J.A., N. Dunbar, D. Rodriguez and SJ Liu (2011) "A Review of Humor in Educational Settings: Four Decades of Research," *Communication Education,* Vol. 60(1), p. 115-144

Bilgen Susanh, Z., and M. Kaytaz (2015) "Determinants of Student Evaluation of Teaching: Evidence from Turkey," *Journal of Business & Economic Policy,* vol. 2(1) p. 121-134

Boring, Ottoboni and Stark (2016) "Student Evaluations of Teaching (mostly) do not Measure Teaching Effectiveness," *Science Open Research,* 11 pages, (DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1)

Bosow, S. A. (1995) "Student Evaluations of College Professors: When Gender Matters." *Journal of Educational Psychology,* vol. 87(4) p. 656-665

Brown, M.J., M. Baillie, and S. Fraser (2009) "Rating Ratemyprofessors.com: A Comparison of Oline and Official Student Evaluations of Teaching," *College Teaching,* Vol. 57(2), p. 89-92

Centra, J. A. and N. B. Gaubatz (2000) "Is there Gender Bias in Student Evaluations of Teaching?," *Journal of Higher Education,* vol. 71(1) p. 17-33

Feldman, K.A. (1984) "Class Size and College Students' Evaluations of Teachers and Courses: A Closer Look," *Research in Higher Education,* Vol. 21(1) p. 45-116.

Garner, R.L. (2006) "Humor in Pedagogy: How Ha-Ha Can Lead to Aha!," *College Teaching,* Vol. 54(1) p. 177-180

Isley, P. and H. Singh (2005) "Do Higher Grades Lead to Favorable Student Evaluations?," *Journal of Economic Education,* Vo. 36(1), p. 29-42.

Jalbert, Terrance (2019a) "Relationships between Business Faculty Teaching and Research Ratings," *Research in Higher Education Journal,* Vol. 36, p. 1-20

Jalbert, Terrance (2019b) "Do Business Professors Who Excel at Both Teaching and Research Exist? *Research in Higher Education Journal,* Vol. 37, p. 1-17

Marks, R.B. (2000), "Determinants of Student Evaluations of Global Measures of Instructor and Course Value, *Journal of Marketing Education,* Vol. 22(2) p. 108-119

Lewis, V.J., and K. McKinzie (2019) "Impact of Industry and Teaching Experience, Course Level and Department on Student Evaluations," *Quarterly Review of Business Disciplines,* Vol. 5(4), p. 335-373

Linsky A.S. and M.A. Strauss (1975) Student Evaluations, Research Productivity and Eminence of College Faculty, *Journal of Higher Education,* Vol. 46(1) p. 89-102

McPherson, M. A. (2006), "Determinants of How Students Evaluate Teachers," *Research in Economic Education,* Vol.37(1, Winter) p. 3-20

Mengel, F., J. Sauermann and U. Zölitz (2018), "Gender Bias in Teaching Evaluations," *Journal of the European Economic Association,* Vol. 17(2), p. 535-566

Miles. P. and D. House (2015) "The Tail Wagging the Dog; An Overdue Examination of Student Teaching Evaluations," *International Journal of Higher Education,* Vol. 4(2), p. 116-126

Miller J. Elizabeth and P. Seldin (2014) "Changing Practices in Faculty Evaluation," *American Association of University Professors, Reports and Publications*, (May-June), downloaded June 25, 2019 from www.aaup.org/article/changing-practices-faculty-evaluation#.XRJN4HdFyUk

Ratemyprofessors.com, accessed June 27, 2019 at: www.ratemyprofessors.com/About.jsp

Sojka, Jane, A.K. Gupta and D.R. Deeter-schmelz (2010) "Student and Faculty Perceptions of Student Evaluations of Teaching: A Study of Similarities and Differences," *College Teaching,* Vol. 50(2) p. 44-49

Wagner, N. M. Rieger and K. Voorvelt (2016) "Gender, Ethnicity and Teaching Evaluations: Evidence from Mixed Teaching Teams," *Economics of Education Review, Vol. 54 (October), p. 79-94*

APPENDIX

Table 1: Number of Responses

Ratemyprofessor.com Tag	Aggregate Responses	Professors with One or More Response	Professors without a Response	Percent with Responses	Mean Number of Responses
Amazing Lectures	287	93	109	46.04	3.086
Hilarious	207	72	130	35.64	2.875
Respected	461	135	67	66.83	3.415
Caring	331	113	89	55.94	2.929
Inspirational	155	75	127	37.13	2.067
Assessible Outside Class	198	95	107	47.03	2.084
Lecture Heavy	374	129	73	63.86	2.899
Participation Matters	247	97	105	48.02	2.546
Skip Class You Won’t Pass	596	138	64	68.32	4.319
Group Projects	163	65	137	32.18	2.508
Get Ready to Read	302	103	99	50.99	2.932
So Many Papers	15	14	188	6.93	1.071
Lots of Homework	379	110	92	54.46	3.446
Test Heavy	261	109	93	53.96	2.395
Clear Grading Criteria	360	128	74	63.37	2.813
Graded by a Few Things	158	83	119	41.09	1.904
Tough Grader	745	157	45	77.72	4.745
Good Feedback	343	128	74	63.37	2.680
Pop Quizzes	78	33	169	16.34	2.364
Extra Credit	180	59	143	29.21	3.051
Would Take Again		108			

This table identifies twenty Tags that students may apply in the process of rating a professor. Students may select up to three Tags for each professor review. Data includes 202 professors and 7,563 individual evaluations. Aggregate Responses indicates the total number of responses received by all sample professors. Professors with one or more response equals the number of professors with at least one response for the Tag. Professors without a response shows the number of professors who did not receive a response for the tag. Percent with responses indicates the percentage of professors that had one or more responses. Mean number of responses specifies the average number of responses for those professors who received the Tag.

Table 2: Single Regression Results

Dependent Var		**Panel A: Raw Evaluations. (N=202)**				**Panel B: Weighted Evaluations. (N=202)**			
Independent Variable	**E.S**	**Intercept**	**Coef.**	**T-Stat**	**R2**	**Intercept**	**Coef.**	**T-Stat**	**R2**
Amazing Lectures	+	3.5573	0.0879	4.88***	0.1064	3.4363	0.0351	3.70***	0.0642
Hilarious	+	3.6221	0.0586	2.98***	0.0424	3.4600	0.0255	2.5**	0.0303
Respected	+	3.5445	0.0603	4.55***	0.0938	3.4331	0.0233	3.34***	0.0527
Caring	+	3.6052	0.0470	3.31***	0.0519	3.4586	0.0168	2.27**	0.0251
Inspirational	+	3.6018	0.1048	3.58***	0.0601	3.4582	0.0365	2.38**	0.0275
Accessible Outside Class	+	3.6015	0.0824	2.85***	0.0390	3.4446	0.0423	2.85**	0.0390
Lecture Heavy	?	3.7076	-0.0137	-1.01	0.0051	3.4990	-0.0069	-0.99	0.0049
Participation Matters	?	3.6618	0.01665	0.63	0.0020	3.4914	-0.0043	-0.32	0.0005
Skip Class You Won’t Pass	?	3.6927	-0.0036	-0.32	0.0005	3.5435	0.0111	1.97*	0.0190
Group Projects	?	3.6993	-0.0213	-0.81	0.0033	3.4889	-0.0035	-0.26	0.0003
Get Ready to Read	-	3.7612	-0.0528	-2.57**	0.0321	3.4953	-0.0062	-0.57	0.0016
So Many Papers	-	3.6907	-0.1154	-0.65	0.0021	3.4866	-0.0058	-0.06	0.0000
Lots of Homework	-	3.6899	-0.0041	-0.27	0.0004	3.4661	0.0107	1.37	0.0093
Course Difficulty	-	5.2069	-0.4634	-6.42***	0.1707				
Test Heavy	?	3.7234	-0.0319	-1.29	0.0082	3.4789	0.0046	0.44	0.0010
Clear Grading Criteria	+	3.6238	0.0327	2.21**	0.0239	3.4761	0.0056	0.73	0.0027
Graded by a Few Things	-	3.7031	-0.0268	-0.81	0.0033	3.5030	-0.0215	-1.26	0.0079
Tough Grader	-	3.7072	-0.0068	-0.85	0.0036	3.4645	0.0059	1.44	0.0102
Good Feedback	+	3.4597	0.0780	3.86***	0.0695	3.4274	0.0346	3.30***	0.0516
Pop Quizzes	?	3.7006	-0.0477	-1.25	0.0078	3.4889	-0.0071	-0.36	0.0007
Extra Credit	+	3.6734	0.0099	0.78	0.0030	3.4837	0.0027	0.42	0.0009
Expertise Area	?	3.5380	0.0997	2.45**	0.0292	3.5407	-0.0378	-1.80*	0.0159
Professor Gender	?	3.632	0.1712	1.56	0.0121	3.5066	-0.0702	-1.24	0.0077
School Quality	+	3.6819	0.0001	0.00	0.0000	3.0857	0.1033	1.25	0.0078
School Reputation	+	2.9410	0.1875	1.43	0.0101	2.8761	0.1543	2.30**	0.0258
School Average Rating	+	2.6839	0.2974	0.37	0.0007	1.7803	0.4570	1.25	0.0077
Public vs Private	?	3.8565	-0.2515	-2.35**	0.0268	3.5710	-0.1224	-2.22**	0.0240
Number of Professors	-	3.7797	-0.0000	-1.23	0.0076	3.5687	-0.0000	-2.04**	0.0205
Dependent Var		**Panel C: Willingness to Retake (N=108)**							
Independent Variable	**E.S.**	**Intercept**	**Coef.**	**T-Stat**	**R2**				
Amazing Lectures	+	66.5734	1.6030	2.31**	0.0479				
Hilarious	+	67.2862	1.5415	2.14**	0.0415				
Respected	+	63.7992	1.8399	3.33***	0.0947				
Caring	+	66.4809	1.3041	2.50**	0.0556				
Inspirational	+	68.1423	1.6635	1.49	0.0204				
Accessible Outside Class	+	65.6756	2.7307	2.49**	0.0552				
Lecture Heavy	?	70.2533	-0.1346	-0.26	0.0006				
Participation Matters	?	67.3451	1.5404	1.47	0.0200				
Skip Class You Won’t Pass	?	70.4735	-0.1240	-0.29	0.0008				
Group Projects	?	70.1099	-0.1922	-0.20	0.0004				
Get Ready to Read	-	73.6551	-1.7340	-2.18**	0.0428				
So Many Papers	-	70.9909	-8.4295	-1.30	0.0158				
Lots of Homework	-	71.1197	-0.4613	-0.80	0.0061				
Course Difficulty	-	105.3084	-10.7697	-2.83***	0.0701				
Test Heavy	?	74.3714	-2.2576	-2.38**	0.0506				
Clear Grading Criteria	+	67.2959	0.09337	1.70*	0.0266				
Graded by a Few Things	-	72.1418	-1.8639	-1.52	0.0213				
Tough Grader	-	72.2589	-0.4228	-1.44	0.0191				
Good Feedback	+	66.7817	1.3096	1.62	0.0243				
Pop Quizzes	?	71.9727	-3.5007	-2.57**	0.0587				
Extra Credit	+	69.1312	0.5177	1.17	0.0128				
Expertise Area	?	66.0349	2.3572	1.12	0.0117				
Professor Gender	?	66.5641	12.0026	2.83**	0.0489				
University Quality	+	31.7040	9.8684	1.31	0.0160				
University Reputation	+	44.8546	6.3730	1.00	0.0094				
University Average Rating	+	-106.3552	47.1721	1.42	0.0187				
Public vs Private	?	66.6786	4.3464	0.81	0.0061				
Number of Professors	-	66.8148	0.0014	0.85	0.0067				

This table shows results of simple regressions. ***, ** and * indicate significance at the 1, 5 and 10 percent levels respectively. E.S. indicates the expected sign.

Table 3: Multiple Regressions

Dependent Variable		**Panel A: Raw Evals.**		**Panel B: Weighted Evals.**		**Panel C: Retake**	
Independent Variable	**E.S.**	**Coefficient**	**T-Stat**	**Coefficient**	**T-Stat**	**Coefficient**	**T-Stat**
Intercept		2.8043		1.1068		-221.0462	
Amazing Lectures	+	0.04763	1.68*	0.03813	2.25**	-0.3205	-0.28
Hilarious	+	0.0087	0.41	-0.0116	-0.93	1.2648	1.55
Respected	+	0.5637	3.00***	0.0189	1.69*	1.9910	2.19**
Caring	+	0.0407	1.58	0.0162	1.05	0.6376	0.62
Inspirational	+	-0.07604	-2.05**	-0.0520	-2.35**	-1.3209	-0.86
Accessible Outside Class	+	0.04978	1.46	0.02936	1.43	2.7579	2.12**
Lecture Heavy	?	-0.0857	-3.98***	-0.0490	-3.81***	-1.844	-2.02**
Participation Matters	?	-0.0253	-1.08	-0.01657	-1.18	0.7197	0.77
Skip Class You Won’t Pass	?	-0.0105	-0.67	0.0045	0.49	0.3703	0.61
Group Projects	?	-0.0394	-1.64	-0.0060	-0.42	0.1516	0.17
Get Ready to Read	-	-0.0328	-1.48	-0.0079	-0.60	-1.6226	-1.90*
So Many Papers	-	-0.0607	-0.41	-0.0496	-0.56	-2.8011	-0.52
Lots of Homework	-	0.0080	0.48	0.0068	0.69	-0.2580	-0.40
Course Difficulty	-	-0.3273	-4.27***			3.8287	0.83
Test Heavy	?	-0.0249	-0.73	0.0020	0.10	-3.5070	-2.75***
Clear Grading Criteria	+	0.05268	2.12**	0.0005	0.03	2.2264	2.42**
Graded by a Few Things	-	-0.0793	-2.06**	-0.0548	-2.39**	-2.2798	-1.55
Tough Grader	-	-0.0086	-0.59	0.0071	0.82	-1.1295	-2.05**
Good Feedback	+	0.0563	2.38**	0.0316	2.23**	-0.5624	-0.58
Pop Quizzes	?	0.0027	0.08	-0.0058	-0.30	-1.4596	-1.21
Extra Credit	+	0.0025	0.12	0.0040	0.26	1.8500	2.33**
Expertise Area	?	0.0602	1.66*	-0.0368	-1.86*	2.4976	1.17
Professor Gender	?	-0.1605	-1.67*	-0.1505	-2.65***	6.4030	1.27
University Quality	+	-0.5544	-1.92*	-0.0950	-0.55	3.7277	0.27
University Reputation	+	0.6119	2.49**	0.2131	1.45	7.9396	0.69
University Average Rating	+	0.4593	0.80	0.5549	1.62	59.4311	2.11**
Public vs Private	?	-0.1451	-1.40	-0.0870	-1.40	5.8166	1.11
Number of Professors	-	-0.0000	-0.10	-0.0000	-1.08	0.0015	1.00
F-Statistic			6.68***		3.38***		4.36***
R2			0.5196		0.3438		0.6071
Adj R2			0.4419		0.2420		0.4679

This table shows results of multiple regressions including all twenty-eight independent variables on the dependent variables. The indicators ***, ** and * indicate significance at the 1, 5 and 10 percent levels respectively. E.S. indicates the expected sign.

Table 4: Stepwise Regression Results

Dependent Variable	**Panel A: Raw Evaluations**			
Variable Entered	**Number of Variables**	**Partial R2**	**Model R2**	**F Value**
Course Difficulty	1	0.1707	0.1707	41.17***
Amazing Lectures	2	0.0880	0.2587	23.61***
Lecture Heavy	3	0.0431	0.3018	12.23***
Caring	4	0.0421	0.3439	12.65***
Public vs Private	5	0.0262	0.3701	8.16***
Graded by a Few Things	6	0.0299	0.4000	9.71***
Accessible Outside Class	7	0.0157	0.4157	5.20**
Get Ready to Read	8	0.0138	0.4295	4.67**
Respected	9	0.0140	0.4435	4.83**
Clear Grading Criteria	10	0.0103	0.4538	3.62*
Inspirational	11	0.0084	0.4622	2.95*
	Panel B: Weighted Evaluations			
Amazing Lectures	1	0.0642	0.0642	13.72***
Graded by a Few Things	2	0.0413	0.1055	9.19***
Accessible Outside Class	3	0.0329	0.1384	7.56***
Public vs Private	4	0.0340	0.1724	8.09***
Lecture Heavy	5	0.0211	0.1935	5.13**
Expertise Area	6	0.0231	0.2166	5.75**
Tough Grader	7	0.0219	0.2385	5.58**
University Average Rating	8	0.0146	0.2531	3.78*
Professor Gender	9	0.0107	0.2638	2.79*
Good Feedback	10	0.0124	0.2762	3.27*
Inspirational	11	0.0114	0.2875	3.03*
Respected	12	0.0121	0.2996	3.29*
University Reputation	13	0.0130	0.3126	3.59*
	Panel C: Willingness to Retake			
Respected	1	0.0947	0.0947	11.09***
Test Heavy	2	0.1370	0.2317	18.73***
Caring	3	0.0757	0.3075	11.37***
Tough Grader	4	0.0350	0.3424	5.48**
Clear Grading Criteria	5	0.0251	0.3675	4.04**
University Quality	6	0.0257	0.3932	4.28**
Lecture Heavy	7	0.0280	0.4212	4.83**
University Average Rating	8	0.0252	0.4464	4.51**
Extra Credit	9	0.0341	0.4804	6.42**
Graded by a Few Things	10	0.0205	0.5010	3.99**
Accessible Outside Class	11	0.0148	0.5158	2.94*
Get Ready to Read	12	0.0173	0.5331	3.52*
Inspirational	13	0.0170	0.5502	3.56*

This table shows stepwise regression results. The approach utilizes forward selection criteria requiring significance at the ten percent level for inclusion in the model. The indicators ***, ** and * indicate significance at the 1, 5 and 10 percent levels respectively.

Table 5: Aggregated Results

Analysis Type		Panel A: Single Regressions			Panel B: Multiple Regressions			Panel C: Stepwise Regressions		
Dependent Variable	**E.S.**	**Raw**	**Weight**	**Retake**	**Raw**	**Weight**	**Retake**	**Raw**	**Weight**	**Retake**
Amazing Lectures	+	+	+	+	+	+		2	1	
Hilarious	+	+	+	+						
Respected	+	+	+	+	+	+	+	9	12	1
Caring	+	+	+	+				4		3
Inspirational	+	+	+		-	-		11	11	13
Assessible Outside Class	+	+	+	+			+	7	3	11
Lecture Heavy	?				-	-	-	3	5	7
Participation Matters	?									
Skip Class You Won't Pass	?		+							
Group Projects	?									
Get Ready to Read	-	-		-			-	8		12
So Many Papers	-									
Lots of Homework	-									
Course Difficulty	-	-		-	-			1		
Test Heavy	?			-			-			2
Clear Grading Criteria	+	+		+	+		+	10		5
Graded by a Few Things	-				-	-		6	2	10
Tough Grader	-						-		7	4
Good Feedback	+	+	+		+	+			10	
Pop Quizzes	?			-						
Extra Credit	+						+			9
Expertise Area	?	+	-		+	-			6	
Professor Gender	?			+	-	-			9	
University Quality	+				-					6
University Reputation	+		+		+				13	
University Average Rating	+						+		8	8
Public vs Private	?	-	-					5	4	
Number of Professors	-		-							

This table shows the combined empirical analysis results. In Panels A and B, a positive notation indicates the variable entered the model significantly with a positive coefficient. A negative notation indicates the variable entered the model significantly with a negative coefficient. Variables without a notation did not enter the model significantly. In Panel C, the cells denote the sequence or stage with which the variable entered the model. E.S. indicates the expected sign.