Course Enrollment Decisions: The Substitutability of Online Ratings Websites for Official Professor Evaluations

Eileen L. Tipoe

May 6 2013

Abstract

Given the growing popularity of professor ratings websites and common departmental policies to keep official professor survey responses confidential, an important concern is the validity of these online ratings. A comparison of student responses to official end-of-semester teaching evaluations and unofficial professor ratings on two widely used websites at UC Berkeley (Rate My Professor and Ninja Courses) indicates that online ratings are significantly lower than their official counterparts. There is also relatively high correlation between official evaluations and online ratings, with most coefficients between 0.4 and 0.7. Similar downward bias was found in other American institutions (Rice University and Harvard University). Some of the bias in Rate My Professor is due to single ratings and early ratings, but similar results are found for Ninja Courses, which has stricter policies regarding posting. Ratings from both websites are not significantly correlated with grade distributions, suggesting that use of these sites for grade retaliation is uncommon. Neither official evaluations nor online ratings are significantly correlated with enrollment.

Acknowledgements

The author would like to thank Professor DellaVigna for his help and valuable input, Emil Schissel for providing the official Economics department data, and various friends who helped with data collection (who prefer to remain anonymous).

1 Introduction

End-of-semester in-class evaluations of professors are collected by departments in many universities worldwide and are used as an indicator of teaching quality, which is often a factor in hiring and promotion decisions. Since students often take into account the difficulty and teaching effectiveness of instructors when selecting the courses to take in future semesters, having this information accessible to students would allow them to make informed choices. Some departments, however, do not publish these official evaluations [1] or show them to students upon request. In those cases students have to rely on unofficial sources such as instructor ratings or schedule planning websites, which have had a growing number of users over this last decade, especially in the past few years (http://www.ratemyprofessors.com/About.jsp).

These sites have a great wealth of information that users have provided about specific aspects of instructors and courses, but there are problems that do not exist with official evaluations. Access to ratings websites is often unrestricted and all comments posted are anonymous, so ratings sites can be used as a form of retribution for bad grades. In comparison, official evaluations are completed before students know their final grade, and so there is little chance that evaluations will be influenced by a students overall performance in the class. Also, users can post multiple comments, and rate the instructor very early on in the semester or without actually having taken the class. There is also some evidence that students use bad ratings as a threat if an instructor does not comply with the students demands, and that instructors even use the site to boost the ratings of themselves or their colleagues, so there have been concerns that these sites are being abused (Brown et al, 2009).

Though extreme cases of abuse have been found to be rare and anecdotal, there are many reasons why ratings may still be biased (Otto et al, 2008). Selection bias is possible, because unlike in-class evaluations, online ratings are completely voluntary, and the students who are willing to take the time to fill in an online survey are likely to be those with strong opinions about instructors. However, the direction of possible bias is ambiguous. Some researchers argue that the students who post ratings and have taken the course for the entire semester are experts due to their extended experience with the professors teaching style, so their advice is valuable to students who plan to take a class taught by that professor (Otto et al, 2008). Those who post may also have discussed the instructor with other students in the class, so the online ratings may represent the collective opinions of a larger group of students than the individual rater. The average online rating of a particular instructor may be much less biased than individual ratings, since highly positive and highly negative ratings could offset each other. Thus, online ratings sites may actually be reliable sources of information.

[1] To distinguish between the official and unofficial sources of data, evaluations refers to official questionnaires distributed by the department that students complete in class, whereas ratings refers to the unofficial surveys.

Given the widespread reliance on unofficial ratings sites for deciding on which instructor to take courses with, it is important to determine whether these unofficial sources are a viable substitute for the official data. The most popular website by far is Rate My Professor, which contains ratings for approximately 1.7 million professors and is used by over 4 million college students across Canada, the United States, and the United Kingdom each month (http://www.ratemyprofessors.com/About.jsp). Making suitable comparisons between ratings from this website and official evaluations can help determine the validity of Rate My Professor, and therefore its usefulness to students. Section 2 discusses existing literature on this topic, Section 3 outlines the data used for this paper, the methodology is outlined in Section 4, Section 5 summarises the findings, Section 6 covers limitations of this study, and Section 7 concludes.

2 Literature Review

Ratings websites have only been established in the past decade, so although there has been extensive research on the traditional in-class evaluations and factors that affect them, only a small number of studies have also incorporated data from Rate My Professor or investigated its validity. However, findings from existing research on Rate My Professor are still very useful for the purposes of this paper.

Previous studies have examined the reliability of either official in-class evaluations or online ratings. While some conduct field experiments to determine how students rate professors, others take a more general approach, using data from several campuses and even different countries. Silva et al (2008) took the latter approach, conducting content analysis of comments about psychology instructors from American and Canadian institutions that students posted on Rate My Professor, in order to determine which characteristics of a course or instructor most frequently result in positive comments. Ratings for approximately 1,100 psychology instructors from 145 institutions in either the US or Canada were coded according to various categories related to aspects of the course, the instructor, and student learning (Silva et al, 2008). Similarly, Otto et al (2008) investigated the link between Rate My Professor ratings and learning outcomes, using data from 373 institutions to test for the presence of a halo effect that affected the accuracy of student reports. This halo effect, found in previous studies of teaching evaluations, is a term used to describe a failure to distinguish between distinct aspects of an individuals behaviour, for example helpfulness and clarity, which results in overly strong correlation between instructor characteristics (Feeley, 2002). The halo effect is stronger when students have strong opinions of the instructor, rate the instructor some time after completing the course, or the evaluation form was not specific in defining categories, so Rate My Professor information could be inaccurate due to this effect (Moritsch and Suter, 1988; Cooper, 1981).

Legg and Wilson (2012) compared the mean responses from a traditional in-class evaluation that a group of college students were required to complete with Rate My Professor ratings that the same group of students were asked to fill in independently, as well as ratings that were on the site before the study. Sonntag et al (2009) also used a relatively small dataset of 126 professors from one college, and examined the correlation between Rate My Professor ratings and official evaluations, as well as easiness with average assigned GPAs of those professors. A similar comparison approach and sample size was used by Brown et al (2009).

The conclusions regarding the validity of Rate My Professor are varied: while some studies believe that Rate My Professor ratings are biased (Legg and Wilson, 2012; Silva et al, 2008), others found that Rate My Professor is a viable alternative for official evaluations (Brown et al, 2009; Sonntag et al, 2009; Otto et al, 2008). Silva et al (2008) find that a significantly greater proportion of comments were positive rather than negative, and focused on aspects of the instructor, such as organization and knowledge, rather than learning outcomes or personal development. These findings do not support the claim that Rate My Professor is primarily a forum for dissatisfied students, and instead suggest that any potential bias in ratings may be in the opposite direction than hypothesized. Studies on official evaluations in earlier decades suggested that instructors who taught more challenging courses or were harsh in assigning grades received lower evaluations in retaliation (Greenwald and Gilmore, 1997; Feldman 1976). Contrary to these findings, recent studies indicate that the effect of student performance on instructor evaluations is also not significant, as students who received high grades did not necessarily give more positive ratings (Centra, 2003). In fact, easiness ratings are thought to have a non-linear relationship, with moderately easy instructors considered as being highly effective, and effectiveness decreasing on either end of the spectrum (Brown et al, 2009). Also, though students factor instructor easiness into their course decisions out of concern for their academic performance, grades were also found to explain only 2 percent of the variability in ratings (Davidson and Price, 2009; Remedios and Lieberman, 2008).

In the Legg and Wilson study, the difference in the mean responses of the Rate My Professor ratings, which the sample was not required to complete, and the compulsory in-class evaluations was statistically significant, with online ratings being lower than the traditional evaluations (Legg and Wilson, 2012). Since previous studies that examined online ratings found that the medium used to complete the survey did not significantly affect the responses as long as students were requested to fill it in (Donovan et al, 2006), the Legg and Wilson data suggests that the self-selected group of students who post on Rate My Professor have a more negative opinion of their instructor. These ratings could have a large impact on the student population: in their preliminary survey, Brown et al (2009) found that 83 percent of their sample use Rate My Professor for information about instructors, and 71 percent have avoided taking a course with a particular instructor because of poor ratings. The heavy reliance on Rate My Professor reflects a high trust level; 96 percent of those surveyed believed that information about instructors from Rate My Professor was as good, if not more accurate than those in official evaluations (Brown et al, 2009). Another study on student perceptions of professors based on online ratings found that reading negative comments establishes prior negative expectations, with a larger response than from reading positive comments (Edwards et al, 2007).

However, Sonntag et al (2009) find that there is high positive correlation between Rate My Professor ratings and official evaluations, and easiness with average assigned GPA, indicating that information on Rate My Professor is similar to that found on official evaluations. Otto et al (2008) concluded that no halo effect exists for Rate My Professor ratings, and that these ratings are indicative of student learning rather than other aspects of the course. Finally, Brown et al (2009) found no significant difference between Rate My Professor ratings and official evaluations, although students gave lower easiness scores and higher helpfulness and clarity scores on official evaluations. Other literature corroborates these results, suggesting that slightly higher official evaluations may be due to social desirability effects i.e. having classmates present while filling in the questionnaire and discussing responses afterward (Ory and Migotsky, 2005; Muhlenfeld, 2005).

One major limitation common to all existing studies is that they do not utilise cross-sectional data. Data for the variables under investigation was either taken from only one semester, or treated as if it was taken from one point in time and separated by professor but not by semester. Only the overall mean responses for ratings categories were used, so any variation in ratings due to the difficulty of a particular course taught by that professor, or improvement in teaching skills over time are not accounted for.

Another issue for some studies was the small sample size: Legg and Wilson (2012) only used 25 students from one university who were all taking the same course, so their sample is far from representative of the student population. Legg and Wilson (2012) also administered the in-class survey early in the semester while the Rate My Professor data was collected at the end of the semester, so the difference in mean responses may be due to more exposure to the professor or improvements or deterioration in teaching style rather than bias. Brown et als survey results that indicate high student confidence in the reliability of information on Rate My Professor also seem less convincing in light of the fact that their sample consisted of only 110 students, most of them psychology majors and college seniors (Brown et al, 2009). Sonntag et al (2009) only sampled professors from one small liberal arts college, so their results may not apply to large universities such as UC Berkeley. Also, the vast majority of papers on professor evaluations and/or ratings are written by those in the field of education or psychology, so this existing body of research would benefit from an economic perspective, particularly the predictive power of online ratings actual decisions such as course enrollment.

A survey of existing literature has indicated that some investigation has already been done to determine the credibility of unofficial professor evaluations, particularly those on Rate My Professor. However, these studies are limited by issues such as small sample size, sampling from only one department or one semester, or using overall professor ratings rather than ratings by semester. Since Rate My Professor is not the only site used by students, it may also be worthwhile to use data from university-specific alternatives to determine whether or not the Rate My Professor findings apply to other sites, in particular those with different user restrictions. This paper hopes to fill the gap in the Rate My Professor-related literature by: 1. Using cross-sectional, micro data at the course and semester level, for each professor. Compared to previous studies that only sampled from one department from one institution or from a non-representative group of less than 200 students, this paper will instead look at four departments with a sample size of thousands of students who filled in these official evaluations. Although most of the analysis will be conducted on the UC Berkeley data, these results will be compared to those at two other institutions (Rice University and Harvard University) to examine external validity. 2. Taking a more detailed approach to Rate My Professor data, in particular, looking beyond mean overall effectiveness ratings or average course grades (which were the only key variables that the large-sample studies used). 3. Making comparisons with ratings from another site (Ninja Courses) with more stringent policies that eliminate some forms of potential bias, an approach that previous studies have not considered.

3 Data

Both official teaching evaluations and unofficial ratings data were collected for 188 undergraduate courses taught by 321 professors from various departments at UC Berkeley from 2006-2012: 63 courses and 119 professors from the Business School, 39 courses and 55 professors from Electrical Engineering (EE) department, 36 courses and 65 professors from the Computer Science (CS) department, and 50 courses and 78 professors from the Economics department. The dataset included all ratings posted on or before the time of collection (March 20, 2013).

3.1 Official Evaluations

The official evaluations were either obtained online from the departmental or department-affiliated websites, or in paper from the department office. Data was not collected from other departments for two main reasons: 1) the department used end-of-semester evaluations solely for the instructors reference and so did not release professor evaluations to students, even under special request, or 2) the data available was too condensed, so comparisons with unofficial data would be severely limited. In some cases, the entire departmental dataset consisted of one summary statistic for overall instructor effectiveness, without even recording the number of respondents.

The 4 departments selected had detailed official datasets containing the number of respondents and the mean response for each question for which students selected a number from a scale. The median and standard deviation for each of these questions were also included, with the exception of the Economics department. To make comparisons between the unofficial sites, the mean response was collected for each question in the official evaluation that most resembled one of the Rate My Professor categories. The official evaluations differed slightly across the 4 departments in terms of wording, but all questionnaires asked students to select an integer from 1-7, or 1-5, to represent how useful the instructors explanations were in understanding the material, how well the instructor identified key points, and how difficult the instructor made the course. These questions were similar to Rate My Professor categories of helpfulness, clarity, and easiness, respectively. The numbers on all scales corresponded to the same extremes so little modification was needed, except for the easiness category of the Econ department data, which was reverse-coded since the highest number on the official questionnaire scale represented the highest level of difficulty, whereas for all other departmental questionnaires the highest number represented the lowest level of difficulty.

Grade distributions for the last 5 years, separated by course and semester, were available on Schedule Builder, the official UC Berkeley website established this academic year to help students plan their course schedules. These distributions consisted of grades submitted by the final grade submission deadline and were not amended after any grade changes in the following semester. Since Schedule Builder did not have distributions for grades submitted after the deadline, Course Rank, another course review and planning website designed for UC Berkeley, was used to obtain these distributions whenever possible. For semesters that both Course Rank and Schedule Builder had grade distributions, the distributions were compared to ensure that Course Rank provided accurate information. There were no instances in which the grade distributions between Course Rank and Schedule Builder differed.

The total number of grades given and number of students that received each letter grade was collected for each semester, and the percentage of students that achieved the letter grades of A+, A, A-, B+, B, B-, C+, C, C-, and D-F was computed from the data, with D-F compiled as a single category. Data for both official evaluations and grade distributions was not available for independent study courses, courses where every student received the same grade, and courses which can only be taken as Pass/Not Pass instead of a letter grade. Grade distributions are also not available for sections with 10 students or fewer. Data from each department was kept and analysed separately, since both official evaluations and unofficial ratings have been shown to differ across subject areas (Felton et al, 2004).

3.2 Unofficial ratings

Data on unofficial ratings was collected from two websites that UC Berkeley students commonly use for information about instructors and courses: Rate My Professor (http://www.ratemyprofessors.com/) and Ninja Courses (ninjacourses.com/). Although the primary focus of this paper is to assess how closely related the Rate My Professor ratings are to the official evaluations, it is also useful to examine whether other sites might be better sources of information about professors and courses. Ninja Courses has some features that may make it a more viable substitute for official evaluations (if unavailable) than Rate My Professor, and these specifics will be discussed later.

3.2.1 Rate My Professor

According to the site description, Rate My Professor is built for college students, by college students, and was intended to help disseminate first-hand information about courses and professors that may not be reflected by the official data or released by departments, such as tips to do well in a particular course, or comments about the professors particular teaching style (http://www.ratemyprofessors.com/About.jsp).

The Rate My Professor survey asks students to rate, on a scale from 1-5, the instructors helpfulness (communication skills and approachability of professor), clarity (ability to understand material based on instructors teaching methods and style), easiness (work required to get a good grade, and how easy it is to get an A in the class), interest level in the class prior to attending, and textbook use (how frequently the professor used the textbook). Each of these categories are clearly explained on a separate page and links to these descriptions embedded in the survey, so there is no ambiguity regarding which aspects of a course or instructor the users are asked to comment on. An individuals comments and responses to all categories except for textbook use are viewable by the public once the survey has been submitted, with the addition of an Overall Quality rating, which is the unweighted average of the helpfulness and clarity ratings. Rate My Professor makes the overall quality rating more salient by displaying a different coloured smiley face for certain ranges of this rating: a yellow smiley face for overall quality ratings of 3.5-5, a green smiley face for ratings of 2.5-3.4, and a blue sad face for ratings of 1-2.4. Another informal survey item is hotness (the professors physical attractiveness), which students select as either hot or not, and if the majority of raters select the former option, a red chili pepper appears beside the professors name. Since the interpretation of hotness is ambiguous and has no analogous category in the official evaluations, this category will not be examined in this study.

The total number of ratings, overall quality rating, and ratings for the other three categories (helpfulness, clarity, and easiness) were collected for each professor. If there was more than one rating for the professor, the number of ratings in each semester was noted, and the individual ratings were used to compute averages of each of the four categories for each semester. The percentage of ratings that were submitted during the first two months of the semester, henceforth referred to as early ratings, was also noted, in order to determine whether users were rating professors before they had had sufficient experience with the course or particular teaching style. Since a rating posted during this time could have been for previous courses that the instructor taught, individual comments were read in cases where it was unclear whether a rating should be classified as early or not. Comments posted early in the semester that referred to final grades received by that individual, multiple midterms that the individual took in that course, or the final exam for that course were assumed to be based on experiences in a past semester and were not classified as early.

3.2.2 Ninja Courses

Ninja Courses is primarily a schedule planning website designed for students on some UC campuses including UC Berkeley, where all lecture/section times, along with the corresponding lecturer are updated for the following semester. Users can rate courses and professors separately, and these ratings are a prominent feature in the course selection menu, with the overall rating for the course and the professor appearing along with the search results. There is also information for each course, including required textbook(s), past instructors (with their ratings included), and users can click on each instructor to see all the individual ratings and comments.

Ninja Courses differs from Rate My Professor in the survey design: students rate professors on a scale of 0-100 on overall quality, difficulty of assignments, difficulty of exams, helpfulness, and easiness. These criteria are not defined in Ninja Courses, but are easily distinguishable and need little clarification. Ratings can only take on integer values, but instead of directly selecting a number from 0-100, each category has a slider that raters can drag along the entire scale, so the ratings might be less precise than those on Rate My Professor. A users ratings of other professors also appear on the scale, so it is easier to rate professors relative to each other. There is a separate slider for the overall quality rating, so it is not a direct function of the other categories. Similar to Rate My Professor, the ratings for each category, along with individual comments, are viewable to other users, and are made more salient using colours: ratings between 70-100 are highlighted in green, ratings between 40-69 are highlighted in orange, and ratings between 0-39 are highlighted in red.

While Rate My Professor accepts ratings for courses that are still in progress, Ninja Courses does not allow users to rate the course or professor until that particular semester has ended. Although this policy does not eliminate the possibility that students use the site as retribution for the grade they receive or that students rate professors or courses they have never taken, it increases the chance that ratings posted are from students who have actually taken the course with that professor for the entire semester. Another useful feature is that users have to select a specific course, semester, and professor to rate, so there is no ambiguity regarding how to classify the data. Finally, unlike Rate My Professor, Ninja Courses does not allow one user to post multiple ratings (unlike Rate My Professor, a user account needs to be created to post on Ninja Courses). The total number of ratings, overall quality, and helpfulness ratings were collected for each professor. Ninja Courses did not have a clarity category. The unweighted average of exam difficulty and assignments difficulty was computed from the Ninja Courses data as the equivalent to Rate My Professors easiness category. Since Ninja Courses used a completely different scale to both the official evaluations and Rate My Professor, data from Ninja Courses was converted to the 0-5 scale. As with the Rate My Professor data, the Ninja Courses ratings for each professor were separated according to the course and semester that it was taught. The average response was computed for each of the categories that were analogous to those on Rate My Professor, with the exception of percentage of early ratings, for reasons discussed in the previous paragraph.

4 Method

In order to examine whether Rate My Professor ratings are a good substitute for official evaluations, the presence of and direction of bias will be determined. There are numerous ways in which the Rate My Professor data may be biased. Firstly, the sample of students that post ratings may not have opinions that are representative of the entire class. To investigate this issue, the average rating for each of the categories will be compared to the corresponding average in the official data, and t-tests for significant differences in the means will be conducted. Early ratings and single ratings may also not be representative of the instructors teaching, and will be compared to the official evaluations to see if this is the case. Finally, students might be rating professors in light of the final grade received, which could result in either higher or lower easiness ratings depending on their grade. The correlation between the easiness category and the mean course grade, as well as the percentage of students that received extremely high or low grades (A- or above, C or below) will be compared for official evaluations, Rate My Professor, and Ninja Courses.

Causal relationships will also be examined. The impact of both official evaluations and online ratings (from both sites) on enrollment will be investigated using a fixed-effects regression, controlling for other variables such as the instructors gender, multiple ratings, and whether or not the course is required for the major.

In order to do correlation and regression analysis, we make the assumption that the ratings data are interval level i.e. the difference between two sequential numbers on the ratings scale is the same as the difference between any other sequential numbers, and a professor with an overall rating of 4 is not twice as good as a professor with an overall rating of 2. Data will also be aggregated by professor, since some observations are missing, particularly in the earlier years.

5 Data Analysis and Findings

5.1 Summary statistics and selection into Rate My Professor

Excerpts of the summary statistic tables for each department[2] are shown in Table 1 below, with the responses for each category from each of the sources (official evaluations, Rate My Professor (RMP), Ninja Courses (NC)) calculated as a weighted average.

The mean responses differ across departments, which may be due to systematic reasons such as the relative difficulty of the courses. Within each department, the mean

[2] See Appendix for the complete summary statistic tables of all variables used

responses for both Rate My Professor and Ninja Courses are all lower than the official evaluations. T-tests were conducted to see whether the difference in means was statistically significant, which was the case in all categories for each department. Of the two ratings sites, Ninja Courses appears to be more widely used in UC Berkeley than Rate My Professor, with a higher number of total ratings for each professor. The total number of ratings per professor on Rate My Professor ranges from 0 to 145 students, which is considerably lower than the total number of students over all semester that have filled in the official evaluations. Thus, it is likely that students who voluntarily fill in online ratings have, on average, more negative opinions about their instructors compared to the average student in that course for that semester.

Table 1: Comparison of Mean Responses (weighted by number of respondents)

CS Department	Official	RMP	NC
OE mean	3.89	3.57**	3.49***
Helpfulness	4.24	3.51***	2.48***
Clarity	4.17	3.60***	-
Easiness	3.85	2.75***	2.33***
N(professor)	65	-	-
Econ Department			
OE mean	3.7	3.33**	3.35***
Helpfulness	3.59	3.36	1.48***
Clarity	3.82	3.31**	-
Easiness	2.81	2.66	1.50***
N(professor)	78	-	-
EE Department			
OE mean	4.14	3.17**	3.59*
Helpfulness	4.45	3.19***	2.22***
Clarity	4.33	3.16***	-
Easiness	3.44	2.93***	2.07***
N(professor)	55	-	-
Business School			
OE mean	4.19	3.56***	3.63***
Helpfulness	4.48	3.45***	2.06***
Clarity	4.34	3.68***	-
Easiness	3.47	3.06**	2.08***
N(professor)	119	-	-

Asterisks indicate the p-value for a t-test for difference in means: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2 (below) shows the Pearson correlation coefficients between official and online ratings, by department, for each category (for example, the correlation between RMP helpfulness and the official evaluation helpfulness rating)[3]. Looking at the results for online ratings, in most cases the easiness coefficient is small and not statistically significant. Strong linear associations were found between the other corresponding categories for the Rate My Professor ratings in the Econ Department, and Ninja Courses ratings in the CS Department. All other cases suggest there is a moderate linear relationship between official and online ratings.

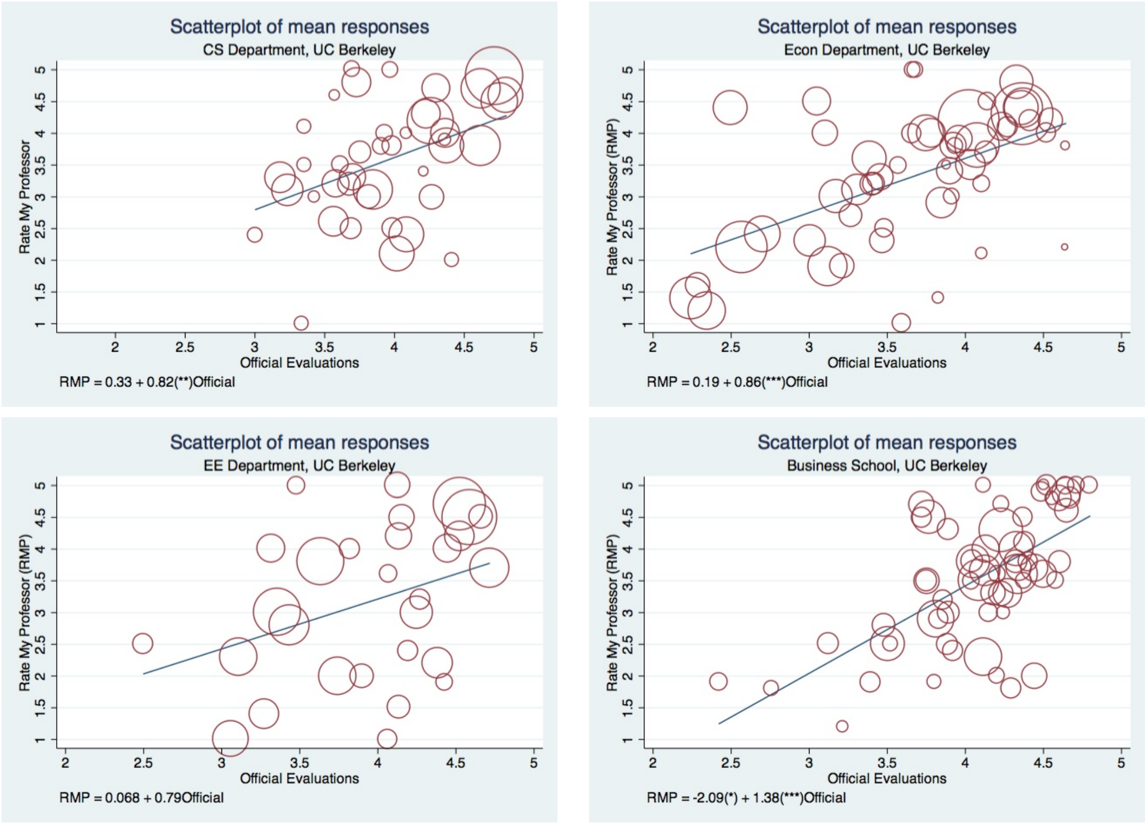
Table 2: Correlation between official and online ratings, by department and professor

CS Department	Official	RMP	NC
Overall effectiveness	-	0.477**	0.826***
Helpfulness	0.852***	0.400*	0.725***
Clarity	0.940***	0.511**	-
Easiness	0.743***	-0.082	0.366*
N(professor)	65		
Econ Department			
Overall effectiveness	-	0.764***	0.654***
Helpfulness	0.993***	0.688***	0.597***
Clarity	0.961***	0.798***	-
Easiness	0.387*	0.288	0.422*
N(professor)	78		
EE Department			
Overall effectiveness	-	0.534*	0.625**
Helpfulness	0.915***	0.551*	0.523*
Clarity	0.934***	0.497*	-
Easiness	0.815***	0.117	0.0852
N(professor)	55		
Business School			
Overall effectiveness	-	0.564***	0.330*
Helpfulness	0.735***	0.558***	0.282
Clarity	0.825***	0.509***	-
Easiness	0.0447	0.172	0.163
N(professor)	119		

Asterisks indicate the p-value for a t-test for difference in means: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[3] Full correlation tables for each department are in the Appendix. Note that Ninja Courses does not have a clarity rating, as mentioned earlier

Figure 1: Scatterplots of overall effectiveness (Official vs Rate My Professor)



Scatterplots of the mean overall effectiveness response for official evaluations (x-axis) and Rate My Professor (y-axis) for each department, weighted by number of respondents, are shown in Figure 1 along with the equation of the fitted line[4] (above):

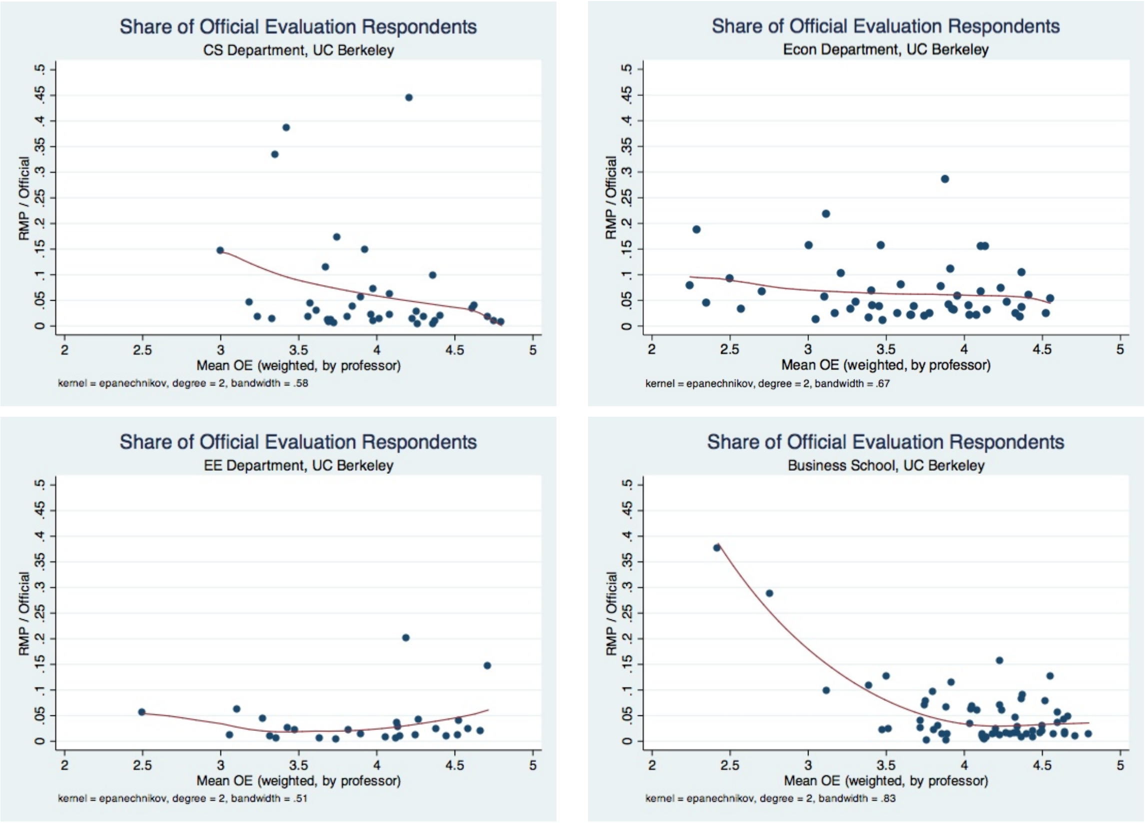
In all cases, there is a positive relationship between the responses on Rate My Professor and those on official evaluations, that is, professors who receive relatively higher official scores generally receive relatively higher RMP ratings. The Rate My Professor ratings have a wider range than the official evaluations; the majority of official evaluation means are in the upper end of the scale (3-5) whereas Rate My Professor ratings are more widely dispersed across the entire scale. Also, there does not seem to be a systematic pattern between number of respondents and the score given, as the size of circles looks unrelated to the numerical rating given. The fitted lines indicate that there is not a one-to-one correspondence between official and unofficial ratings, with positive y-intercepts and statistically significant slope coefficients less than 1 in most cases, so professors rated officially as average receive 'below-average' Rate My Professor ratings.

[4] Asterisks beside the coefficients in the equation indicate the p-value for a t-test for difference in means: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

It would be useful to see whether students select into Rate My Professor based on the quality of their instructor. We would expect that the negative bias in online ratings is due to a larger proportion of students posting for professors with lower teaching quality. Scatterplots of the share of students that post on Rate My Professor as a proportion of official survey respondents, plotted against the mean overall effectiveness (by professor, from the official data) are shown above in Figure 2, with a fitted polynomial.

For all departments, the students that rate online make up a very small proportion of those who filled in the official evaluation form (less than 1 percent, although there were some outliers with values between 1-10 percent that were not shown here, for graph comparison purposes). There is a slight negative relationship between quality of professor and share of online raters, mostly clearly seen in the CS department graphs, indicating that professors with relatively more online ratings were officially evaluated as average. In other departments there is no clear evidence that selection into Rate My Professor varies with instructor quality. However, due to little variation in share of respondents, there is no evidence to suggest that this selection effect into Rate My Professor is large. The fitted polynomial would be more accurate if there were observations for professors with ratings of 1-2.5, but (perhaps fortunately) such ratings do not exist in the data.

Figure 2: Relationship between share of respondents and mean official evaluation response



As mentioned earlier, Ninja Courses may be less biased due to features of the site that prevent one user from submitting multiple ratings for the same professor, and rating the professor before the current semester is over. The same scatterplots were made for the Ninja Courses data, and are shown below. Here, we find similar patterns to Rate My Professor. Professors are generally rated lower online than on the official evaluations, but there is little evidence to suggest that selection into Ninja Courses varies with instructor quality.

Figure 3: Scatterplots of overall effectiveness (Official vs Ninja Courses)

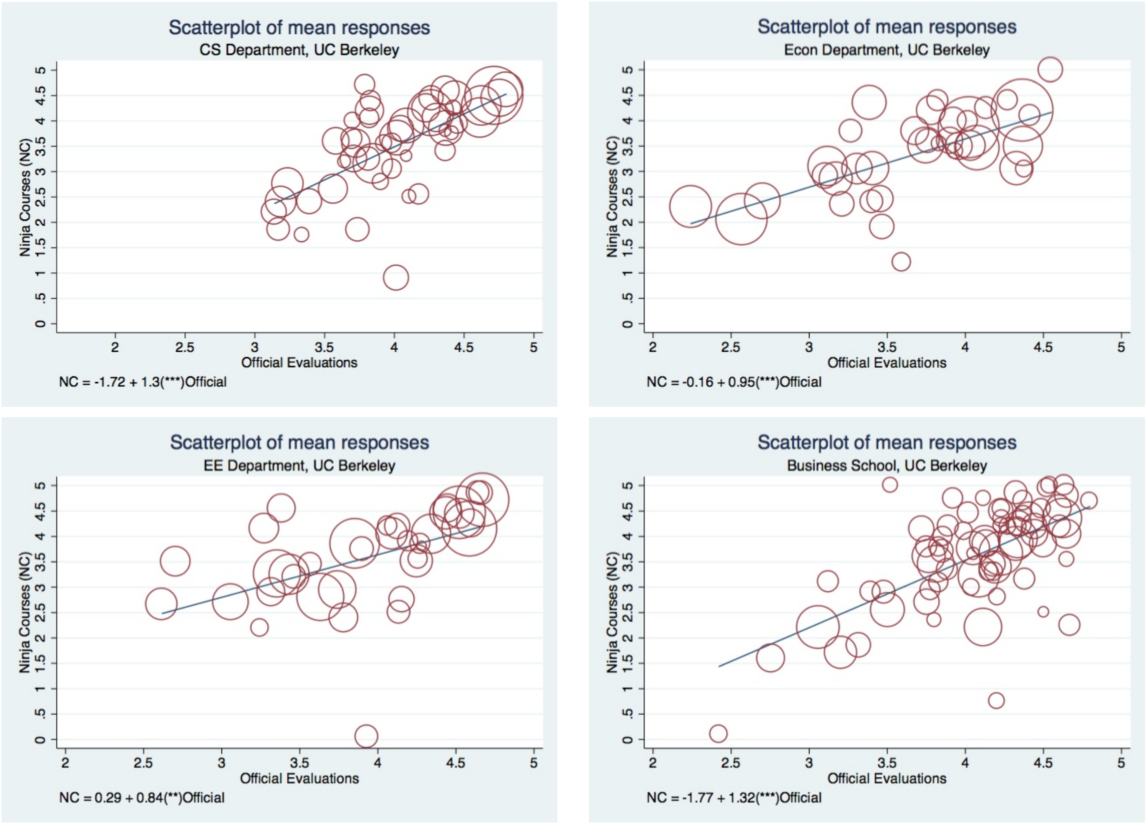
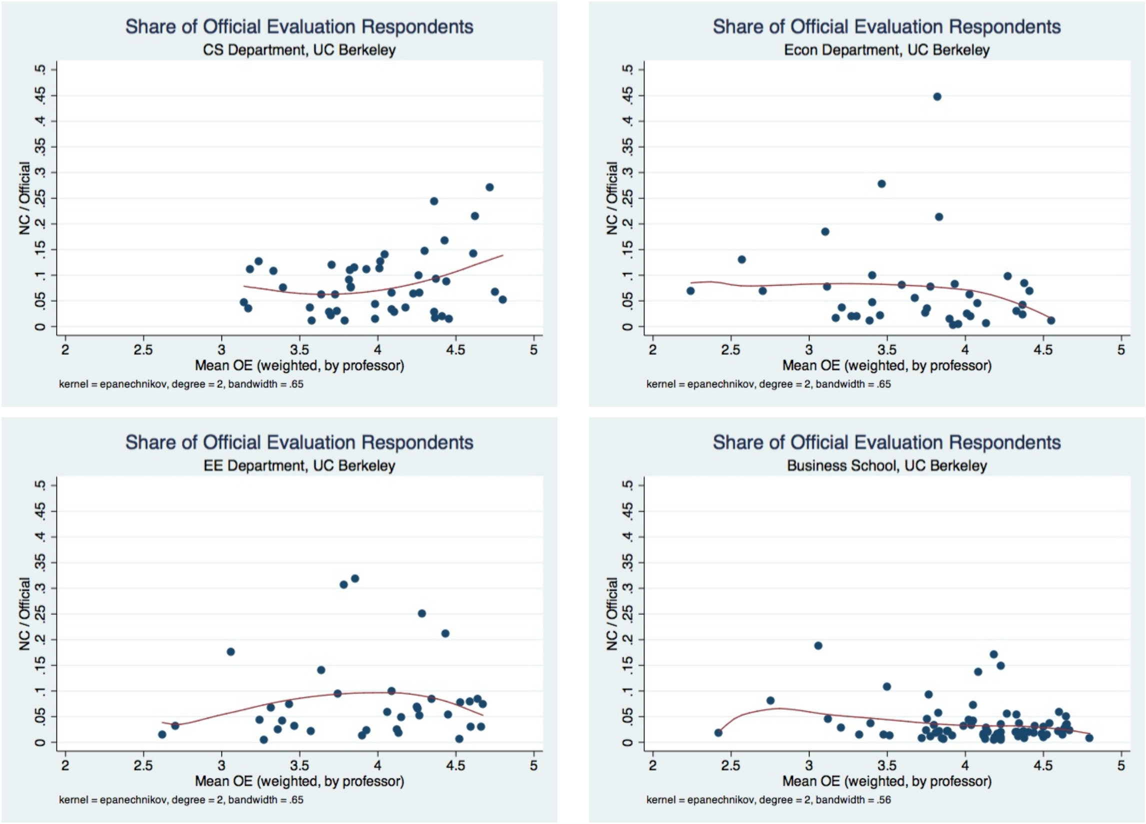


Figure 4: Relationship between share of respondents and mean official evaluation response (Ninja Courses)



5.1.1 Comparison to other institutions

Are these findings similar to those in other institutions? Data was collected for the corresponding departments in Rice University, as well as the CS Department in Harvard University[5]. These were the only institutions where it was possible to gain access to the data without a university user account. The same graphs as in Figures 1 and 2 were plotted and are shown below, along with a similar table for mean responses. As in the Berkeley data, Rice University professor ratings are, on average, lower than the corresponding official evaluations, though there are some outliers. Harvard University has a slight bias in the other direction, though there may be too few observations to discern a pattern and neither of the coefficients for the fitted line is statistically significant. Similar to the Berkeley finding, the selection graphs show weak evidence of selection into Rate My Professor, with no distinct pattern to support the hypothesis that the share of Rate My Professor respondents is higher if instruction quality is lower.

Table 3: Comparison of Mean Responses (weighted by number of respondents)

CS Department (Rice)	Official	RMP
OE mean	4.09	3.46
Helpfulness	4.15	3.66
Clarity	3.98	3.23
Easiness	2.78	2.89
N(professor)	73	-
Econ Department (Rice)		
OE mean	3.79	2.67***
Helpfulness	4.11	2.56***
Clarity	3.7	2.82***
Easiness	2.83	2.34*
N(professor)	73	-
EE Department (Rice)		
OE mean	4.09	3.52*
Helpfulness	4.25	3.60**
Clarity	4.02	3.45*
Easiness	2.72	2.56
N(professor)	108	-
CS Department (Harvard)		
OE mean	4.25	3.77*
N(professor)	14	-

[5] See Appendix for summary statistics tables for all departments collected. Note that Rice University only offers Business as a minor and there were very few classes offered, so the Business Department was omitted from this analysis. Rice University official data was also reverse coded so the 1-5 scale corresponded to that of Rate My Professor

Figure 5: Relationship between share of respondents and mean official evaluation response (Rice University and Harvard University)

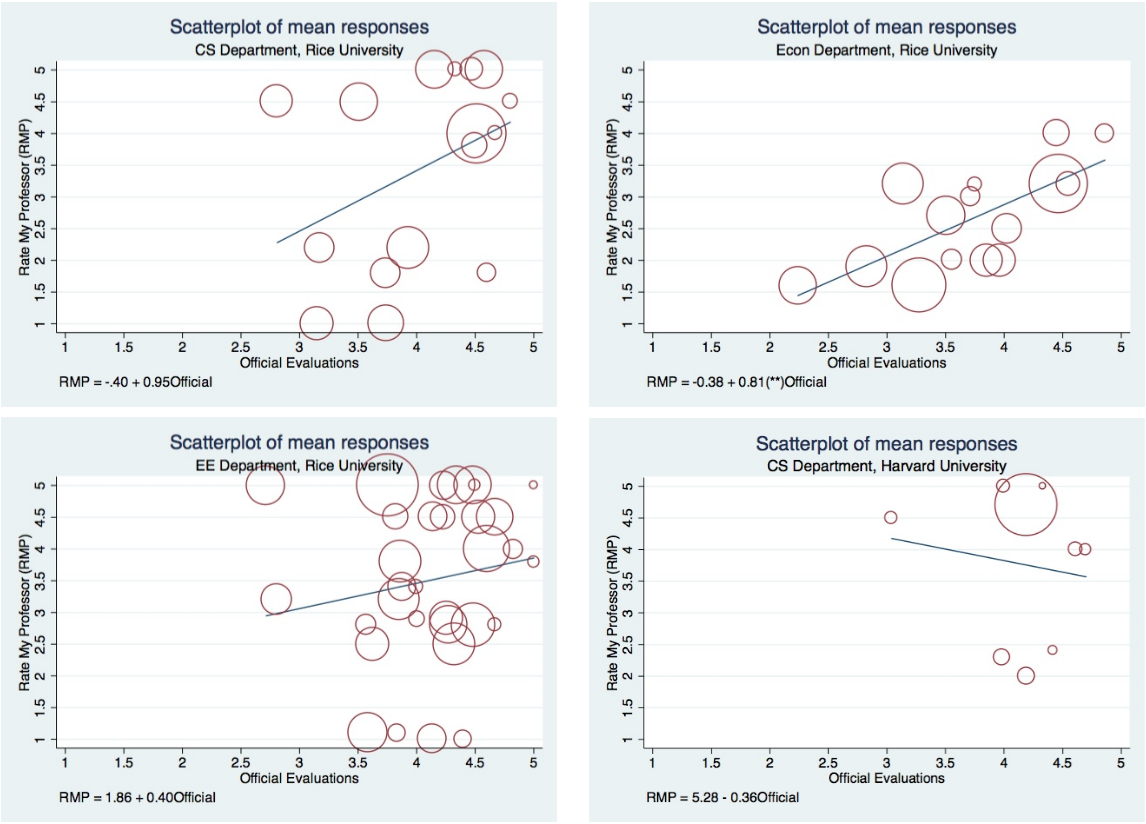
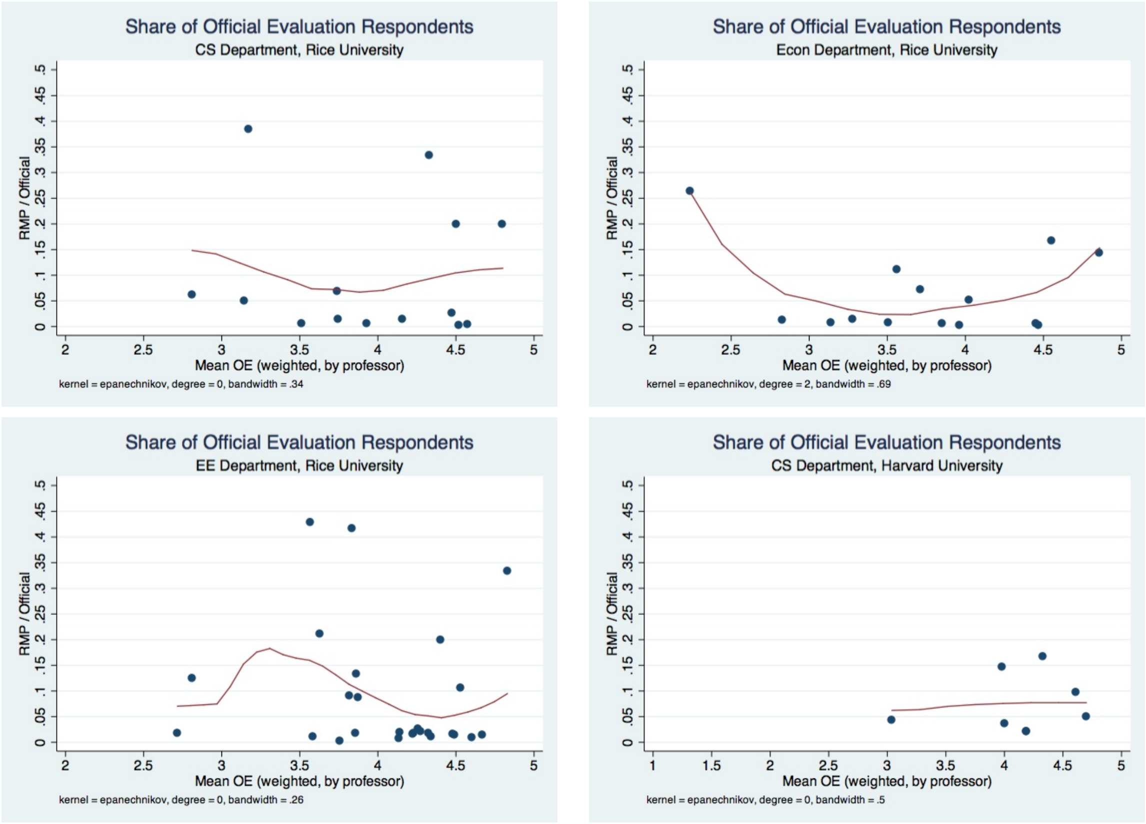


Figure 6: Relationship between share of respondents and mean official evaluation response (Rice University and Harvard University)



5.2 Bias due to features of Rate My Professor

Some studies have postulated that major sources of bias in Rate My Professor aside from a non-representative sample are students who rate too early in the semester due to strong initial impressions of the professor, and the fact that a sizeable number of professors only have one rating (Brown et al, 2009; Otto et al, 2008). To test whether or not these biases are present, the mean response for each category in Rate My Professor was obtained with single ratings and early ratings omitted separately. These responses, along with official evaluations and overall Rate My Professor ratings, are summarized in Table 2 below[6]:

Table 4: Single and early ratings from Rate My Professor, by department

CS Department	Official (Total)	RMP (Total)	Official (single)	RMP (single)	Official (early)	RMP (early)
OE mean	3.89	3.57	4.07	3.79**	3.88	3.34*
Helpfulness	4.24	3.51	4.37	3.74***	4.26	3.23***
Clarity	4.17	3.6	4.29	3.81***	4.15	3.44**
Easiness	3.85	2.75	3.96	2.75***	3.86	2.54***
% of total RMP ratings	-	-	-	11.89	-	9.69
Econ Department						
OE mean	3.8	3.44	3.6	3.35***	3.64	3.54
Helpfulness	3.87	3.45	3.51	3.34**	3.54	3.54
Clarity	3.87	3.44	3.7	3.36***	3.75	3.55*
Easiness	2.82	2.68	2.78	2.65*	2.81	2.79
% of total RMP ratings	-	-	-	3.4	-	10.68
EE Department						
OE mean	3.97	3.16	4.05	3.51***	3.96	3.23*
Helpfulness	4.21	3.15	4.25	3.45***	4.24	3.26***
Clarity	4.29	3.17	4.36	3.56***	4.3	3.19*
Easiness	3.93	2.76	3.9	2.98***	3.85	3.06***
% of total RMP ratings	-	-	-	37.58	-	6.06
Business School						
OE mean	4.19	3.56	4.08	3.54***	4.06	3.65***
Helpfulness	4.48	3.36	4.42	3.47***	4.38	3.51***
Clarity	4.34	3.77	4.28	3.62***	4.26	3.79**
Easiness	3.47	3.26	3.48	3.01**	3.39	3.27
% of total RMP ratings	-	-	-	14.36	-	9.82

Asterisks indicate the p-value for a t-test for difference in means: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In most cases, the mean response with single or early online ratings omitted was higher that the total mean online response. These results indicate that students who rate professors early on in the semester tend to have a more negative opinion of the professor than the average student in the class, and that single ratings are also from the more dissatisfied students in the course. For all departments except Economics, downward bias still exists: Rate My Professor ratings were still lower than the corresponding official evaluations, and the difference in means across all categories was still statistically significant at the 5% level. In most departments, early and single ratings are only a small

[6] See Appendix for the full summary statistic tables by department. Asterisks indicate the p-value for a t-test for difference in means: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

proportion of total Rate My Professor ratings (15% or less of total ratings) and so cannot entirely account for the bias found in Section 5.1.

5.3 'Easiness' and grade distributions

Previous literature has postulated that grade retaliation could happen when posting on anonymous websites but has not examined how prevalent this issue is. Official evaluations are filled in before students know their final grades, whereas Rate My Professor and Ninja Courses are usually completed after final grades are submitted (a required condition for Ninja Course postings). According to the hypothesis that students do use online sites for retaliation, correlation between easiness and unofficial ratings is expected to be stronger than in official evaluations. The Pearson correlation coefficients of the easiness category of each of the sources and the grade distributions (by department, per course per semester), along with the mean course grade, are summarized in Table 3 below:

Table 5: Pearson correlation coefficients (easiness and grade distributions), by professor

CS Department	Official	RMP	NC
% A- and above	0.0476	0.252	-0.0244
% C and below	0.175	-0.169	0.188
Mean grade	-0.0747	0.247	-0.0918
N(professor)	65		
Econ Department			
% A- and above	0.0746	-0.0311	-0.038
% C and below	-0.333*	0.131	0.072
Mean grade	0.203	-0.0703	-0.104
N(professor)	78		
EE Department			
% A- and above	0.289	-0.0404	0.193
% C and below	0.109	0.131	-0.101
Mean grade	0.106	-0.00342	0.0826
N(professor)	55		
Business School			
% A- and above	-0.262	0.283*	0.0487
% C and below	0.432**	-0.23	0.0788
Mean grade	-0.419**	0.243	0.0793
N(professor)	119		

Asterisks indicate the p-value for a t-test for difference in means: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The correlation between easiness and final grades assigned is as expected: professors who give a larger percentage of students As or a higher average grade tend to get higher ratings, and those who give a larger percentage of students Cs and/or failing grades tend to get lower ratings. One exception is the statistically significant positive coefficients corresponding to the lowest grades in the Business School, which may be because students who rated those professors were not the same students who received those grades. However, in all cases, the correlation coefficients are small, with absolute value of 0.5 or less, indicating a weak linear relationship. Most of the coefficients are also not significantly different from zero, so it is important not to interpret these results literally. There are also no systematic differences between the coefficients for Rate My Professor/Ninja Courses ratings and the official evaluations, though the official evaluation correlation coefficients are generally larger than the Ninja Courses coefficients.

There are many possible reasons for the lack of a strong relationship between easiness and final grades. Students may be more concerned with the amount they learnt from the course rather than the grade they achieved, and therefore base their ratings heavily on the professors teaching style unless the professors grading policy is extremely unfair or lenient. A data-related issue is that ratings are anonymous, making it impossible to determine what grade the rater received in that class. Inability to match raters to the grades they received could explain results that are contrary to the hypothesis, such as the statistically significant positive coefficients corresponding to the lowest grades in the Business School. Students might base their ratings on the grade they received relative to others rather than the absolute grade assigned, but this effect cannot be measured due to anonymity of raters. Finally, since easiness is relative and more open to interpretation than the other categories, there may be some correlation between grades and easiness that is dependent on the raters aptitude for that course but cannot be found using the data.

The assumed linearity between these variables in question may not hold: existing literature has suggested that easiness and grades have a non-linear relationship, but this hypothesis has not been tested (Brown et al, 2009). The Spearman correlation coefficients of the easiness category of each of the sources and the grade distributions (by department, per course per semester) are summarized in Table 4 on the following page.

The Spearman correlation coefficients are slightly larger than the Pearson correlation coefficients in most cases but there is no particular pattern for these differences, so it is not certain that easiness and grades have a non-linear but positive relationship. Some of the statistically significant coefficients are for the extreme higher grades, suggesting that high ratings are motivated by good grades, but again the inability to match raters to their grades may be the reason why some coefficients are the opposite sign than expected. Statistically significant correlation coefficients were only found for the official data, so students may be reasonably good at predicting the grade they will receive in the course and evaluating the professor accordingly, and there is no evidence that online ratings are based on grades received. It is also possible that the relationship between easiness ratings and grades is not monotonic, so neither correlation coefficient can accurately capture the patterns in the data.

Table 6: Spearman correlation coefficients (easiness and grade distributions), by professor

CS Department	Official	RMP	NC
% A- and above	0.0918	0.218	0.0374
% C and below	0.157	-0.196	0.102
Mean grade	-0.0734	0.269	-0.00603
N(professor)	65		
Econ Department			
% A- and above	0.0593	0.00517	-0.02
% C and below	-0.359*	0.151	0.0333
Mean grade	0.232	-0.0546	-0.0568
N(professor)	78		
EE Department			
% A- and above	0.319	-0.0817	0.131
% C and below	0.0441	0.133	-0.134
Mean grade	0.142	-0.13	0.0309
N(professor)	55		
Business School			
% A- and above	-0.247	0.266	0.121
% C and below	0.405**	-0.138	-0.0152
Mean grade	-0.370**	0.211	0.16
N(professor)	119		

Asterisks indicate the p-value for a t-test for difference in means: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5.4 Correlation between ratings and enrollment

How are these rating patterns related to actual enrollment decisions? To investigate this question, the percentage enrollment of courses in these departments for the Spring 2013 semester were calculated as a decimal (between 0 and 1) from the UC Berkeley Class Schedule page (http://schedule.berkeley.edu). This variable was regressed separately on the mean overall effectiveness response on official evaluations and Ninja Courses using a fixed effects regression (by course), controlling for easiness, gender of the instructor, whether or not there were multiple ratings for the instructor, and whether or not the course was required for that major. The Rate My Professor regression omitted the other categories because the overall effectiveness category was a simple average of helpfulness and clarity, and would result in a multicollinearity problem. Also, the Business School was omitted from this analysis because enrollment in all undergraduate courses was either 99 or 100 percent. Under the assumption that students use professor ratings (either official or unofficial) from previous semesters to make course decisions, the impact of ratings on enrollment can be determined. These regressions are summarized in Table 5 on the following page.

Contrary to the hypothesis that students largely base enrollment decisions on some form of rating, the coefficients for overall effectiveness and easiness are not statistically significant, except for Ninja courses ratings for Economics and Computer Science Departments. The negative coefficient for the Econ Department can be explained by the constant, which is larger than 1. These results are somewhat surprising, especially for online ratings, given the large number of users and the bias in ratings, but may be due to the small number of observations since courses in the dataset are not offered every semester; this issue will be discussed further in Section 6. The reported intercept, which is the mean of the fixed effects, is statistically significant in most cases, so course-specific characteristics do correlate with enrollment. The coefficients for gender are also statistically significant, with less students enrolling for classes taught by female professors. This result is consistent with findings from earlier studies (Baldwin and Blattner, 2003). Aside from the EE Department, multiple ratings do not significantly correlate enrollment, possibly because students either do not depend on online ratings for those decisions, or because students take the ratings at face value without considering the number of ratings. Statistically significant coefficients of a large magnitude were found for required class, meaning that fulfilling major requirements may be a bigger concern for students than the professor teaching that class.

This regression may not present a full picture of the course decision process for many reasons. Some students may not have an option to wait for a more highly rated professor if they wish to graduate in a timely manner, so ratings would not be the deciding factor in course decisions. Also, the regression could have omitted variables that have a greater influence on enrollment but were either difficult to measure, or data is unavailable. Some examples are: time of the class (students may avoid early morning or late-night classes), time of the final exam (generally the last slot on Friday is the least popular), or taking classes with friends (the benefits of which could compensate for lower quality of instruction). These characteristics, along with other professor-specific characteristics, were assumed to be time-invariant in order to run the fixed-effects regression but may not be since teaching quality could improve over time and exam/class schedules could vary every semester, so fixed effects does not perfectly model the situation.

Table 7: Fixed-effects regression results (% enrollment and explanatory variables)

	Computer Science (% enrollment)			Economics(% enrollment)			Electrical Engineering (% enrollment)		
	Official	RMP	NC	Official	RMP	NC	Official	RMP	NC
Constant	0.802***	2.651	-0.526	0.095***	0.702**	1.190***	0.267	0.541*	0.501*
	(-15.32)	(-1.32)	(-3.15)	(-8.61)	(-3.85)	(-9.02)	(-1.28)	(-2.66)	(-2.69)
Overall effectiveness	-0.0142	-0.985	0.920***	-1.29	0.0202	-0.175*	0.0663	0.00426	-0.063
	(-0.48)	(-1.66)	(-5.4)	(-0.35)	(-0.64)	(-2.54)	(-0.79)	(-0.16)	(-0.87)
Easiness	0.0449	0.646	0.891***	2.39	0	0.0964	0.000302	-0.08919	-0.0157
	(-1.58)	(-1.89)	(-6.48)	(-0.3)	(.)	(-1.49)	(.)	(-0.72)	(-0.13)
Female	0	0	0	-0.152***	-0.0703*	-0.0545	-1.33	0	-0.0748
	(.)	(.)	(.)	(-15.80)	(-2.58)	(-1.71)	(-1.07)	(.)	(-0.64)
Required class	0	0	-0.00992	0	0	0	0.199*	0.221**	0.190*
	(.)	(.)	(-0.15)	(.)	(.)	(.)	(-2.51)	(-2.87)	(-2.13)
Multiple ratings	-	0	-0.202	-	0.0356	0	-	0.297*	0
	-	(.)	(-1.23)	-	(-0.6)	(.)	-	(-2.74)	(.)
N(course total ratings)	17	-	-	13	-	-	11	-	-

Asterisks indicate the p-value for a t-test for difference in means: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6 Limitations

6.1 The data

As mentioned in Section 3, only a small number of departments kept detailed records of student responses to official end-of-semester evaluations, so it would be impossible to determine whether the same patterns apply to other departments at UC Berkeley. However, this study improved on the approach of previous studies by incorporating cross-sectional data from some of the largest departments on campus and so accounts for a substantial percentage of the student population within one university. There was considerable heterogeneity in the data, since these departments had a large number of faculty members as well as a wide range of courses and class sizes, and so was reasonably representative of the entire population.

Not all courses in each department were offered every semester, and since professors in the dataset were responsible for teaching only one or two different courses for the entire time period considered, some professors taught their course(s) more than others. Although official evaluations were aggregated so that each professor had more observations for Rate My Professor and Ninja Courses, experience in lecturing and writing exams for a particular course can affect both teaching ratings and grade distributions. Having enough unofficial data to conduct the same analysis as in Section 5 would be necessary to examine the trends in ratings and evaluations over time. In some departments, two or more professors taught the course in the same semester and were evaluated separately but had the same grade distributions, so it is hard to tell the effect of a particular professors easiness on ratings or evaluations. There were also many cases in which the grade distributions were quite skewed with the lowest grade being a B or B-, especially in small classes, so it is hard to separate the students ability from the professors generosity in grading. However, there are few statistically significant correlations between easiness and grade distributions, and easiness only makes a small contribution to enrollment, so this issue is not major. The regression only included percentage enrollment for the Spring 2013 semester, since this information was not available for earlier semesters and could not be inferred from the data because maximum class size differs every semester. To improve the quality of the regression, this study could be repeated once enrollment data from future semesters is available.

The wording of the questionnaire or categories affects student responses. Although the official surveys are generally quite clearly worded, for unofficial sites the precise definitions of categories are not made salient at the time of rating. In the case of Rate My Professor, the definitions helpfulness and clarity are slightly counterintuitive and overlap. For example, when rating an instructors helpfulness, a student should consider whether the professor explains clearly. Rate My Professor provides a link to this information, but it takes more effort to ensure that ratings comply with the sites definitions compared to official questionnaires and so students may resort to relying on their own interpretation of categories to save time. If in addition students cannot clearly distinguish between categories, the overall rating could be based on one category (the halo effect as discussed in Section 2). Another issue is whether raters can convey accurate information to other users by their comments; it is worth considering possible differences in interpretation between writing and reading ratings. Investigating these issues would shed light on possible sources of bias and may suggest ways to improve the reliability of ratings sites.

6.2 The methodology and data analysis

A major flaw in Rate My Professor is that the site allows one user to submit multiple ratings for the same professor. The anonymous nature of the comments meant that the prevalence of multiple comments posted by the same user on Rate My Professor could not be investigated. Attempts to identify possible instances of multiple ratings were severely hampered by the fact that these ratings may not have been posted at the same time, and most comments were too brief to determine whether the language or content was similar. Rate My Professor adheres to privacy regulations and will therefore not release student information, even to researchers, so it is unlikely that this issue can be examined in the future.

Content analysis using criteria similar to Silva et al (2008) using data from various departments rather than solely psychology instructor ratings was not conducted for this study due to time and budget constraints, but would help answer the research question. Having a large proportion of online ratings giving constructive feedback about their courses and instructors would mitigate the negative bias in these ratings, and support the conclusion that Rate My Professor is a viable alternative for official evaluations, especially since official evaluations do not contain all student comments made.

There are large amounts of data available at the undergraduate level due to the larger class sizes, but little is available at the graduate level. Graduate programs, especially at the doctoral level, are usually more focused on research than courses, so the quality of a professors teaching might be a much smaller concern to a graduate student than the professors research interests. As an extension of this study, it would be interesting to see whether ratings and evaluations still have the same relationship when the group of students rating may have different priorities when selecting instructors, as well as less interest in the rating system itself.

The external validity of this study should be the focus of future research. This paper examined whether the Rate My Professor ratings have a similar relationship to the official data for other universities, including those that do not have alternative unofficial sources like Berkeley does. However, are there any country-specific differences in the official vs. unofficial ratings? Ninja Courses is specific to UC Berkeley and some other UC campuses; do the Ninja Courses findings from this paper also apply to those universities? These questions can be answered using official evaluation data from some of the other 8,000 or so institutions from the US, Canada, and the UK that are also on Rate My Professor.

7 Conclusion

Analysis of official evaluations and unofficial ratings from four departments at UC Berkeley indicates that online ratings are, on average, significantly lower than their official counterparts. There is also relatively high correlation between official evaluations and online ratings from both sites, with most coefficients between 0.4 and 0.7. There is little evidence that students are selecting into Rate My Professor based on the quality of their instructor, with a similar share of students rating online for professors that received lower official evaluation scores compared to professor with higher official scores. These patterns were further investigated by examining official evaluations and ratings from the same departments at peer institutions (Rice University and Harvard University) where there are no alternative sites such as Ninja Courses. The findings were similar to those found in the Berkeley data, with significantly lower mean responses for each category in most departments but no evidence of selection into Rate My Professor.

Some of the bias is due to single ratings as well as user regulations on Rate My Professor that allow students to submit ratings early on in the semester; omitting these from the dataset result in higher means for each category but still significant differences with the official responses. Early ratings is not a complete explanation since Ninja Courses, a site that does not allow such ratings, has downward bias similar in magnitude to Rate My Professor. However, there is no evidence of grade retaliation on either of the ratings websites, as there is no meaningful significant correlation between grade distributions of a particular professor and his/her easiness score. A large part of the existing bias still needs to be explained, perhaps by examining how students behaviour or attitudes may differ depending on the nature of the survey, or investigating how wording of the questionnaire may affect the quality of responses.

Online ratings are not significantly related to student enrollment within these departments so this bias may not have serious economic consequences for the Berkeley student population, though future studies could modify the regression to include hard-to-measure variables that are important in course enrollment decisions. Nevertheless, since these websites are often the only source of information about professors available to students, measures should be taken to ensure that online ratings are representative of the opinions expressed in official evaluations.

8 References

Baldwin, T., and N. Blattner. 2003. Guarding against potential bias in student evaluations: what every faculty member needs to know. College Teaching 51, no. 1: 2732.)

Brown, M.J., M. Baillie, and S. Fraser. 2009. Rating RateMyProfessors.com: A comparison of online and official student evaluations of teaching. College Teaching 57: 8992.)

Centra, J.A. 2003. Will teachers receive higher student ratings by giving higher grades and less course work? Research in Higher Education 44: 495518.)

Cooper, W. H. (1981). Ubiquitous halo. Psychological Bulletin, 90, 218 244.)

Course Rank. Course Rank. 2012. 26 03 2013 <https://www.courserank.com/berkeley/main>.

Davidson, E., and J. Price. 2009. How do we rate? An evaluation of online student evaluations. Assessment and Evaluation in Higher Education 34: 5165

Donovan, J., C.E. Mader, and J. Shinsky. 2006. Constructive student feedback: Online vs. traditional course evaluations. Journal of Interactive Online Learning 5: 28396.

Edwards, C., A. Edwards, Q. Qing, and S.T. Wahl. 2007. The influence of computer-mediated word-of-mouth communication on student perceptions of instructors and attitudes toward learning course content. Communication Education 56: 25577.)

Eta Kappa Nu, University of California Berkeley. Course Surveys. 12 2012. 15 02 2013 <https://hkn.eecs.berkeley.edu/coursesurveys/instructors>.

Feeley, T.H. 2002. Evidence of Halo Effects in Student Evaluations of Communication Instruction, Communication Education, 51:3, 225-236

Feldman, K.A. 1976. Grades and college students evaluations of their courses and teachers. Research in Higher Education 4: 69111

Felton, J., Mitchell, J., and Stinson, M. (2004). Web-based student evaluation of professors: The relations between perceived quality, easiness and sexiness. Assessment and Evaluation in Higher Education, 29, 91108.)

Greenwald, A.G., and G.M. Gilmore. 1997. Grading leniency is a removable contaminant of student ratings. American Psychologist 52: 120917

Haas School of Business, University of California Berkeley. Student Evaluations of Faculty. 12 2012. 30 01 2013 <https://aai.haas.berkeley.edu/TIES/>.

Harvard University. The Q Evaluations. 2013. 01 05 2013 <http://q.fas.harvard.edu/harvardQ/shopping.jsp>.

Legg, A. M. and Wilson, J. H. (2012): RateMyProfessors.com offers biased evaluations, Assessment and Evaluation in Higher Education, 37:1, 89-97

Li, William and Alex Sydell. Ninja Courses. 2013. 05 03 2013 <http://ninjacourses.com/>.

Moritsch, B. G., and Suter, W. N. (1988). Correlates of halo error in teacher evaluation. Educational Research Quarterly, 12, 29 34.)

Mhlenfeld, H. 2005. Differences between talking about and admitting sensitive behaviour in anonymous and non- anonymous web-based interviews. Computers in Human Behavior 21:9931003.

Otto, J., D.A. Sanford Jr., and D.N. Ross. 2008. Does ratemyprofessor.com really rate my professor? Assessment and Evaluation in Higher Education 33: 35568.

Ory, J. C., and C. P. Migotsky. 2005. Getting the most out of your student ratings of instruction. APS Observer
<http://www.psychologicalscience.org/observer/19/7/teaching_tips/>.

Rate My Professor, LLC. Rate My Professors. 2013. 01 03 2013
<http://www.ratemyprofessors.com/>.

Remedios, R., and D.A. Lieberman. 2008. I liked your course because you taught me well: The influence of grades, workload, expectations and goals on student evaluations of teaching. British Educational Research Journal 34: 91115

Rice University. Office of the Registrar. 16 04 2013. 30 04 2013
<http://registrar.rice.edu/students/evals/>.

Silva, K. M., Silva, F. J., Quinn, M. A., Draper, J. N., Cover, K. R., Munoff, A. A. 2008. Rate My Professor: Online Evaluations of Psychology Instructors, Teaching of Psychology, 35:2, 71-80

Sonntag, M. E., Bassett, J. F., Snyder, T. 2009. An empirical test of the validity of student evaluations of teaching made on RateMyProfessors.com, Assessment and Evaluation in Higher Education, 34:5, 499-504

University of California Berkeley. Schedule Builder. 2012. 26 03 2013
<https://schedulebuilder.berkeley.edu/>.

9 Appendices

Table 8: Summary statistics (collapsed weighted means, by department)

	Computer Science				Economics				Electrical Engineering				Business School			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Number of respondents	63.37	57.76	9	347	76.73	79.16	4	433.1	40.88	32.72	4	138.6	48.18	38.8	9	194.6
Percentage enrollment	0.9	0.11	0.7	1	0.86	0.18	0.14	0.99	0.65	0.18	0.4	0.97	-	-	-	-
OE mean	3.89	0.52	1.6	4.8	0.8	0.22	0.1	1	0.66	0.19	0.4	1	4.14	0.46	2.4	4.9
Helpfulness	4.24	0.48	2.2	4.9	3.59	0.61	2	4.7	4.22	0.53	2.4	5	4.45	0.32	3.4	4.9
Clarity	4.17	0.4	2.3	4.8	3.82	0.47	2.6	4.6	4.27	0.4	3.3	5	4.33	0.32	3.3	4.9
Easiness	3.85	0.36	2.1	4.5	2.81	0.21	2.2	3.3	3.91	0.45	2.8	5	3.44	0.39	2.6	4.3
RMP ratings per prof.	6.82	12.65	0	68	12.91	22.71	0	125	3.2	7.58	0	50	5.4	15.16	0	145
OQ (RMP)	3.57	0.92	1	5	3.33	1.03	1	5	3.17	1.23	1	5	3.56	1.02	1.2	5
Helpfulness (RMP)	3.51	0.97	1	5	3.36	1.01	1	5	3.19	1.3	1	5	3.45	1.13	1	5
Clarity (RMP)	3.6	0.94	1	5	3.31	1.1	1	5	3.16	1.19	1	5	3.68	1.03	1.1	5
Easiness (RMP)	2.75	0.86	1	5	2.66	0.76	1	5	2.93	0.72	1	4	3.06	0.96	1	5
Early ratings (%)	0.13	0.22	0	1	0.08	0.14	0	0.5	0.37	0.77	0	3.3	0.21	0.34	0	1
NC total ratings	27.46	52.35	0	253	11.49	28.45	0	142	11.22	24.42	0	165	11.13	39.33	0	331
OQ (NC)	3.49	0.88	0.9	4.7	3.35	0.83	1.2	5	3.59	0.95	0.1	4.8	3.63	1.01	0.1	5
Helpfulness (NC)	2.48	1.65	0	4.6	1.48	1.66	0	5	2.22	1.75	0	4.9	2.06	1.87	0	5
Easiness (NC)	2.33	1.53	0	4.3	1.5	1.64	0	4.4	2.07	1.62	0	5	2.08	1.82	0	5
Female	0.05	0.21	0	1	1	0	1	1	1	0	1	1	1	0	1	1
Required course	0.45	0.5	0	1	0.3	0.43	0	1	0.24	0.36	0	1	0.35	0.45	0	1
N(professor)	65	-	-	-	78	-	-	-	55	-	-	-	119	-	-	-

Table 9: Correlation between official and online ratings (CS Department)

	OE	Helpfulness	Clarity	Easiness	OE(RMP)	Helpfulness(RMP)	Clarity(RMP)	Easiness(RMP)	OE(NC)	Helpfulness(NC)	Easiness(NC)
OE	1										
Helpfulness	0.852***	1									
Clarity	0.940***	0.769***	1								
Easiness	0.743***	0.599***	0.737***	1							
OE(RMP)	0.477**	0.319	0.469**	0.570***	1						
Helpfulness(RMP)	0.400*	0.255	0.416*	0.559**	0.965***	1					
Clarity(RMP)	0.511**	0.354	0.475**	0.536**	0.966***	0.866***	1				
Easiness(RMP)	-0.082	-0.265	-0.0682	0.0837	0.18	0.2	0.147	1			
OE(NC)	0.826***	0.750***	0.767***	0.663***	0.575***	0.499**	0.612***	-0.0484	1		
Helpfulness(NC)	0.725***	0.724***	0.667***	0.647***	0.580***	0.538**	0.589***	-0.0257	0.909***	1	
Easiness(NC)	0.366*	0.256	0.415*	0.502**	0.598***	0.614***	0.552**	-0.0324	0.612***	0.577***	1

Table 10: Correlation between official and online ratings (Econ Department)

	OE	Helpfulness	Clarity	Easiness	OE(RMP)	Helpfulness(RMP)	Clarity(RMP)	Easiness(RMP)	OE(NC)	Helpfulness(NC)	Easiness(NC)
OE	1										
Helpfulness	0.993***	1									
Clarity	0.961***	0.956***	1								
Easiness	0.387*	0.382*	0.431*	1							
OE(RMP)	0.764***	0.738***	0.650***	0.2	1						
Helpfulness(RMP)	0.688***	0.647***	0.579***	0.158	0.975***	1					
Clarity(RMP)	0.798***	0.789***	0.686***	0.227	0.979***	0.913***	1				
Easiness(RMP)	0.288	0.26	0.281	0.532**	0.554**	0.577***	0.511**	1			
OE(NC)	0.654***	0.658***	0.550**	0.0905	0.742***	0.707***	0.734***	0.246	1		
Helpfulness(NC)	0.597***	0.590***	0.488**	0.000712	0.718***	0.708***	0.695***	0.223	0.892***	1	
Easiness(NC)	0.422*	0.425*	0.366*	0.278	0.571***	0.551**	0.558**	0.450*	0.763***	0.694***	1

Table 11: Correlation between official and online ratings (EE Department)

	OE	Helpfulness	Clarity	Easiness	OE(RMP)	Helpfulness(RMP)	Clarity(RMP)	Easiness(RMP)	OE(NC)	Helpfulness(NC)	Easiness(NC)
OE	1										
Helpfulness	0.915***	1									
Clarity	0.934***	0.826***	1								
Easiness	0.815***	0.740***	0.807***	1							
OE(RMP)	0.534*	0.482*	0.498*	0.311	1						
Helpfulness(RMP)	0.551*	0.528*	0.508*	0.395	0.977***	1					
Clarity(RMP)	0.497*	0.417	0.475*	0.227	0.974***	0.903***	1				
Easiness(RMP)	0.117	0.248	0.214	0.0855	0.543*	0.606**	0.452*	1			
OE(NC)	0.625**	0.519*	0.642**	0.722***	0.354	0.371	0.336	-0.0175	1		
Helpfulness(NC)	0.523*	0.513*	0.486*	0.513*	0.359	0.34	0.377	-0.0406	0.840***	1	
Easiness(NC)	0.0852	0.0114	0.183	0.355	0.0738	0.103	0.0479	0.012	0.624**	0.381	1

Table 12: Correlation between official and online ratings (Business School)

	OE	Helpfulness	Clarity	Easiness	OE(RMP)	Helpfulness(RMP)	Clarity(RMP)	Easiness(RMP)	OE(NC)	Helpfulness(NC)	Easiness(NC)
OE	1										
Helpfulness	0.735***	1									
Clarity	0.825***	0.844***	1								
Easiness	0.0447	-0.095	-0.00533	1							
OE(RMP)	0.564***	0.569***	0.574***	-0.000352	1						
Helpfulness(RMP)	0.558***	0.546***	0.572***	0.074	0.957***	1					
Clarity(RMP)	0.509***	0.541***	0.519***	-0.0949	0.940***	0.804***	1				
Easiness(RMP)	0.172	0.275	0.307*	-0.535***	0.449**	0.388**	0.478***	1			
OE(NC)	0.330*	0.268	0.447**	-0.278	0.492***	0.460**	0.488***	0.275	1		
Helpfulness(NC)	0.282	0.261	0.426**	-0.343*	0.391**	0.376**	0.378**	0.244	0.915***	1	
Easiness(NC)	0.163	0.0547	0.295*	-0.343*	0.306*	0.320*	0.272	0.349*	0.814***	0.815***	1

Table 13: Summary statistics (early ratings omitted)

	Computer Science				Economics				Electrical Engineering				Business School			
Official	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Overall Quality	3.88	0.44	1.6	4.5	3.64	0.59	2.3	4.6	3.96	0.57	2.6	4.8	4.06	0.47	2.8	4.9
Helpfulness	4.26	0.41	2.2	4.8	3.54	0.64	2	4.7	4.24	0.53	2.4	5	4.38	0.38	3.4	4.9
Clarity	4.15	0.33	2.3	4.7	3.75	0.47	2.7	4.6	4.3	0.4	3.3	5	4.26	0.34	3.3	4.9
Easiness	3.86	0.3	2.1	4.4	2.81	0.21	2.2	3.3	3.85	0.51	2.8	5	3.39	0.42	2.6	4.3
Rate My Professor	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Overall Quality	3.34	0.92	1	5	3.54	1.03	1	5	3.23	1.14	1	5	3.65	0.77	1.8	5
Helpfulness	3.23	0.98	1	5	3.54	1.01	1	5	3.26	1.21	1	5	3.51	0.89	1	5
Clarity	3.44	0.95	1	5	3.55	1.11	1	5	3.19	1.13	1	5	3.79	0.8	1.8	5
Easiness	2.54	0.96	1	5	2.79	0.74	1	5	3.06	0.61	1	4	3.27	0.89	1.5	5
N(course total ratings)	275	-	-	-	219	-	-	-	239	-	-	-	500	-	-	-

Table 14: Summary statistics (single ratings omitted)

	Computer Science				Economics				Electrical Engineering				Business School			
Official	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Overall Quality	4.07	0.52	1.6	4.8	3.6	0.66	2.2	4.6	40.88	32.72	4	138.6	4.08	0.48	2.4	4.9
Helpfulness	4.37	0.45	2.2	4.9	3.51	0.74	2	4.7	3.94	0.57	2.5	4.8	4.42	0.35	3.4	4.9
Clarity	4.29	0.38	2.3	4.8	3.7	0.53	2.6	4.6	0.66	0.19	0.4	1	4.28	0.33	3.3	4.9
Easiness	3.96	0.31	2.1	4.5	2.78	0.2	2.2	3.3	4.22	0.53	2.4	5	3.48	0.4	2.6	4.3
Rate My Professor	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Overall Quality	3.79	0.87	2.1	4.9	3.35	1.04	1	4.8	3.51	1.06	1.4	4.7	3.54	0.86	1.2	5
Helpfulness	3.74	0.89	1.8	5	3.34	1	1	4.7	3.45	1.03	1.5	4.7	3.47	0.93	1.3	5
Clarity	3.81	0.85	2.3	4.9	3.36	1.11	1	5	3.56	1.09	1.3	4.8	3.62	0.87	1.1	5
Easiness	2.75	0.74	1.2	5	2.65	0.62	1.4	4.2	2.98	0.5	1.7	3.7	3.01	0.94	1	5
N(course total ratings)	268	-	-	-	237	-	-	-	159	-	-	-	474	-	-	-

Table 15: Summary statistics for Rice University and Harvard University (weighted, by professor)

	Computer Science				Economics				Electrical Engineering				Computer Science (Harvard)			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Number of respondents	16.39	12.84	1	59.7	26.09	19.42	1.5	86	16.48	13.47	1	66.7	79.71	165.66	1	643
OE mean	4.09	0.52	2.6	5	3.79	0.6	2.2	5	4.09	0.52	2	5	4.25	0.48	3	5
Helpfulness	4.15	0.49	2.6	5	4.11	0.49	2.8	5	4.25	0.45	2.5	5				
Clarity	3.98	0.61	2.2	5	3.7	0.61	2.2	5	4.02	0.55	2.4	5				
Easiness	2.78	0.66	1	4.2	2.83	0.38	2.1	4	2.72	0.65	1.1	4.2				
RMP ratings per prof.	2.47	2.48	0	8	3.07	3.08	1	10	2.59	1.43	1	5	4	4.56	1	14
OQ (RMP)	3.46	1.51	1	5	2.67	0.8	1.6	4	3.52	1.27	1	5	3.77	1.21	2	5
Helpfulness (RMP)	3.66	1.61	1	5	2.56	1.06	1	4	3.6	1.34	1	5				
Clarity (RMP)	3.23	1.55	1	5	2.82	0.72	1.8	4	3.45	1.33	1	5				
Easiness (RMP)	2.89	1.17	1	5	2.34	0.99	1	4	2.56	1.05	1	4				
N(professor)	73	-	-	-	73	-	-	-	108	-	-	-	14	-	-	-