Correlations, trends, and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors.com data

Andrew S. Rosen

Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA

Corresponding author: Andrew S. Rosen, rosen@u.northwestern.edu

Andrew S. Rosen is a graduate student in the Department of Chemical and Biological Engineering at Northwestern University. He holds a B.S. in Chemical Engineering from Tufts University (2015). His main research interests are in molecular modelling, with additional broad interests in data science.

Abstract. Student evaluations of teaching are widely adopted across academic institutions, but there are many underlying trends and biases that can influence their interpretation. Publicly accessible web-based student evaluations of teaching are of particular relevance due to their widespread use by students in the course selection process and in the quantity of data available for analysis. In this study, data from the most popular of these websites, RateMyProfessors.com, is analysed for correlations between measures of instruction quality, easiness, physical attractiveness, discipline, and gender. This study of 7,882,980 RateMyProfessors ratings (from 190,006 U.S. professors with at least 20 student ratings) provides further insight into student perceptions of academic instruction and possible variables in student evaluations. Positive correlations were observed between ratings of instruction quality and easiness as well as between instruction quality and attractiveness. On average, professors in science and engineering disciplines have lower ratings than in the humanities and arts. When looking at RateMyProfessors as a whole, the effect of a professor’s gender on rating criteria is small but statistically significant. When analysing the data as a function of discipline, however, the effects of gender are significantly more pronounced, albeit more complex. The potential implications of the aforementioned trends are discussed.

Keywords: student evaluations of teaching; RateMyProfessors; online evaluations; rating correlations; gender bias
Introduction

Student evaluations of teaching

Student evaluations of teaching are a major component of academic instruction at colleges and universities. These evaluations are typically administered by the institution and completed by students anonymously (Abrami, D’Apollonia, and Rosenfield 2007). Some institutions use well-defined procedures to administer these evaluations in the classroom (Cashin 1995), while others rely on the completion of web-based surveys (Ballantyne 2003). In addition to being used as a way to continuously improve teaching and course quality based on student feedback, scores from student evaluations have potential implications for academic faculty through means such as promotion, tenure, and awards (Abrami, d’Apollonia, and Rosenfield 2007; Benton and Cashin 2014). Despite their widespread use, the effective design, administration, and interpretation of student evaluations to properly account for student biases is a topic of much debate (Marsh and Bailey 1993). As a result, it is essential to understand the underlying correlations and trends among various criteria in student evaluations of teaching. The popularity of online, publicly accessible faculty review websites makes it possible to analyse some of these trends in an automated and systematic fashion on a large scale.

Overview of RateMyProfessors.com

Launched in 1999, RateMyProfessors.com is a review website owned by mtvU that is dedicated to anonymous, user-submitted reviews of college and university professors. As of December 2016, RateMyProfessors has approximately 17 million ratings for over 1.6 million professors at nearly 7,000 institutions, making it the largest publicly available database of student-submitted teaching evaluations at institutions of higher education (“About RateMyProfessors.com” 2016). Both formally administered student evaluations of teaching and websites such as RateMyProfessors are a significant part of the conversation about course selection across many colleges and universities (Steffes and Burgee 2009; Bleske-Rechek and Fritsch 2011). In one study of RateMyProfessors use, 92% of the 216 surveyed students had heard of the site, 80% had visited the website more than once, 84% of students found the site helpful, and 95% found it credible (Davison and Price 2009). In addition, it has been shown that RateMyProfessors may impact student expectations of potential classes (Kowai-Bell et al. 2011) as well as perceptions of the instructor once enrolled (Lewandowski, Higgins, and Nardone 2012). The large sample size of ratings and the popularity of RateMyProfessors make it an insightful dataset in the context of web-based student evaluations of teaching at institutions of higher education and is therefore the focus of this study.

Students on RateMyProfessors rate professors based on three main key criteria: clarity, helpfulness, and easiness. These criteria are mandatory fields when a review is submitted, and the rating scale ranges from 1 to 5 in integer increments with 1 being the worst rating and 5 being the best rating. Professors are then assigned an overall quality score that is simply the average of their clarity and helpfulness scores. For visual presentation purposes, RateMyProfessors breaks down overall quality into three categories: poor (1.0–2.4), average (2.5–3.4), and good (3.5–5.0).

In this work, the clarity, helpfulness, and overall quality scores are collectively referred to as metrics of ‘instruction quality’ to distinguish them from the easiness
Therefore, account for treatment of hotness again scores that included both positive and negative Stins points higher than prof found that attractive professors had student ratings study, (hotness sexiness aspects of the person in question person influences the reviewer's ability to distinguish among conceptually distinct the halo effect, a type of confirmation bias instruction quality \( r \) (ratings hotness score divided by the number of submitted reviews) but now with a proper data set that is greater than 0, the professor is considered 'hot', and an image of a chili pepper is displayed on the professor's RateMyProfessors profile.

It should be noted that as of 18 May 2016, RateMyProfessors switched to a more simplistic rating system and omitted the clarity and helpfulness scores from the website entirely. Students are now asked to explicitly rate the overall quality of the professor in the submission form. This change removed two categories that provide insight into professor ratings and in turn modifies the interpretation of the overall quality metric. As such, the dataset considered in this study is limited to submissions collected prior to the 18 May 2016 change. While this change removes valuable metrics from the RateMyProfessors site, it is also possible that it will increase the validity of the overall quality metric in measuring teaching effectiveness since most institutional student evaluations of teaching do not judge overall instruction quality based solely on an instructor's clarity and helpfulness.

**Related research**

A number of researchers have analysed various aspects of the publicly available RateMyProfessors dataset. Felton, Mitchell, and Stinson (2004) examined the relationship between instruction quality, easiness, and sexiness (defined as the hotness score divided by the number of submitted reviews for a given professor, where the hotness score had a minimum value of zero). For 1,148 faculty with at least 20 student ratings, the correlation between overall quality and easiness was found to be \( r = 0.67 \) \((p < 0.01)\), and the correlation between overall quality and sexiness was found to be \( r = 0.31 \) \((p < 0.01)\). The authors also postulated that the relationship between instruction quality and easiness for professors with high sexiness scores is indicative of the halo effect, a type of confirmation bias where a reviewer's overall impression of a person influences the reviewer's ability to distinguish among conceptually distinct aspects of the person in question (Thorndike 1920). However, their definition of sexiness was incomplete, as it did not take into account professors with negative total hotness scores and also did not acknowledge that hotness is an optional rating metric (i.e. only some students rate whether their professors are ‘hot’ or ‘not hot’). In a related study, Riniolo et al. (2006) analysed the effect of attractiveness on RateMyProfessors student ratings based on 522 professors at the four schools with the most ratings and found that attractive professors had overall quality scores that were approximately 0.8 points higher than professors rated as less attractive.

Felton et al. (2008) provided an update to the work of Felton, Mitchell, and Stinson (2004) by considering a larger dataset of 6,582 faculty members and hotness scores that included both positive and negative values. The results of this work once again found statistically significant correlations between overall quality, easiness, and hotness. In their work, hotness was evaluated in the same way as sexiness (i.e. the hotness score divided by the number of submitted reviews) but now with a proper treatment of hotness scores less than zero. As previously stated, this definition does not account for the number of times a professor was rated on their physical attractiveness. Therefore, this quantitative value cannot be interpreted as the percent of reviews that
indicate a professor is ‘hot’ (for positive weighted hotness scores) or ‘not hot’ (for negative weighted hotness scores) due to the optional nature of the question.

In order to address one of the most essential questions regarding the evaluation of RateMyProfessors data, Otto, Sanford, and Ross (2008) sought to answer whether RateMyProfessors really does ‘rate my professor’. Their conclusion, based on the online ratings of 399 randomly selected faculty members from 373 institutions, emphasizes that, while there are certainly plausible cases where online rating sites like RateMyProfessors can be abused, the online ratings from RateMyProfessors ‘showed a similarity with what might be expected if the ratings were valid measures of student learning’ (Otto, Sanford, and Ross 2008). This was demonstrated via a strong positive correlation between clarity and helpfulness as well as a variability in easiness that was inversely proportional to clarity and helpfulness. This led the authors to claim that RateMyProfessors ratings were not subject to the halo effect as previously reported by Felton, Mitchell, and Stinson (2004) since the variance in easiness scores was not similar for all levels of clarity and helpfulness, indicating that students were able to distinguish between the rating criteria.

The aforementioned work is supported by Bleske-Rechek and Michels (2010) who report several trends based on student reviews of 322 professors at the University of Wisconsin–Eau Claire on RateMyProfessors. The authors conclude that written reviews are ‘wide ranging and moderate in tone’ (as opposed to either ranting or raving), there are few academic characteristics (e.g. GPA, learning goal orientations) that differentiate those who review professors on RateMyProfessors and those who do not, and discipline-related differences in easiness and overall quality indicate that students effectively differentiate a professor’s easiness and quality of instruction.

Nevertheless, recent research has questioned these findings. The work of Legg and Wilson (2012), for instance, seems to contradict Bleske-Rechek and Michels (2010) by suggesting that students that leave evaluations on RateMyProfessors have a negative bias compared with formal in-class evaluations, although this analysis only included 25 faculty members across a number of departments at a single university. Further, a study by Clayson (2014) refuted the claim made by Otto, Sanford, and Ross (2008) and instead suggests that RateMyProfessors is an invalid measure of teaching effectiveness if effectiveness is interpreted as teaching ability. Instead, Clayson states that RateMyProfessors – as well as many institutional student evaluations of teaching – are most likely reflective of a ‘likeability’ scale.

While there is much debate about the validity of RateMyProfessors in measuring instruction quality, it has been shown that there is a strong association between RateMyProfessors scores and institutional student evaluations of teaching. In a study by Coladarci and Kornfield (2007), the results of RateMyProfessors ratings for professors at the University of Maine were compared with formal in-class student evaluations. The results of this study based on 283 instructors reveal a strong correlation between the RateMyProfessors overall quality scores and the evaluation item ‘Overall, how would you rate the instructor?’ ($r = 0.68, p < 0.001$). The authors found a slightly weaker correlation between RateMyProfessors easiness scores and the evaluation item ‘How did the work load for this course compare to that to others of equal credit?’ ($r = 0.44, p < 0.001$). Sonntag, Bassett, and Snyder (2009) found that the RateMyProfessors easiness scores of 126 professors at Lander University were positively correlated ($r = 0.44, p < 0.01$) with the end-of-term class grades. Similarly, clarity and helpfulness scores were positively correlated with institutionally administered evaluation items related to overall excellence of instruction ($r = 0.69, p <$
0.01) and overall excellence ratings of the class itself ($r = 0.60, p < 0.01$). The authors also replicated prior results by Greenwald and Gillmore (1997) and Felton, Mitchell, and Stinson (2004) that reflect a bias toward grading leniency in both traditional and publicly available web-based student evaluations of teaching.

Motivation

While there is a strong association between RateMyProfessors scores and those of institutional student evaluations of teaching, the validity of RateMyProfessors in measuring teaching effectiveness is still a topic of much debate. As discussed by Clayson (2014), even if RateMyProfessors is not a valid metric of teaching effectiveness, it is still worth serious consideration, as this would imply that institutional student evaluations of teaching contain potentially invalid metrics as well (at least to the degree that they are associated with the scores on RateMyProfessors). As such, if there are trends and biases in the RateMyProfessors data, there is a high likelihood that there are similar differences in institutional student evaluations. A better understanding of publicly accessible student evaluations of teaching such as those available on RateMyProfessors therefore has the ability to influence the interpretation and implementation of evaluations of teaching at the college and university level. Collectively, the motivation of this work is to shed light on trends that are likely applicable to both public and institutionally administered student evaluations of teaching.

By analysing nearly 13.7 million ratings from web-based student evaluations of teaching, this work serves to validate and expand upon the findings of prior research related to RateMyProfessors that is often performed for significantly smaller – and potentially unrepresentative – datasets (e.g. at a single institution or for a single discipline at a small number of schools). The specific trends in the RateMyProfessors rating criteria are investigated not only in terms of the overall correlations but also the complex ways in which various criteria are dependent on one another. In addition, this work more properly accounts for the optional attractiveness rating on the site that prior research on RateMyProfessors has neglected.

The trends in rating criteria are discussed in the context of a number of parameters relevant to institutions of higher education. By comparing the scores of various disciplines, potential differences in the ratings of professors in science, technology, engineering, and mathematics (STEM) fields compared to those in the humanities and arts can be observed. The influence of physical attractiveness and course difficulty on professor ratings is discussed as well.

Furthermore, this is the first study to look at the effects of professor gender on web-based student evaluations of teaching on such a large scale. Since RateMyProfessors does not include professor gender on its site, this work utilizes an automated method based on historical data to reliably predict professor gender. This enables an unprecedented look into the effects of professor gender on student evaluation scores both on the site as a whole and on a discipline-by-discipline basis.

Methodology

The raw RateMyProfessors dataset used in this study is based on 13,677,171 ratings of 1,231,643 professors from 4,522 colleges and universities within the United States. This represents all of the publicly available reviews on RateMyProfessors from the fifty states (and Washington, D.C.) as of 3 January 2016, which is just four months before
the clarity and helpfulness scores were removed from the site (and, by extension, before the overall quality score was redefined). This study is therefore the most comprehensive RateMyProfessors dataset that has been studied in the literature.

The majority of the data analysis was performed using Python 3 with the aid of the pandas data structure library (McKinney 2010), the NumPy scientific computing library (van der Walt, Colbert, and Varoquaux 2011), the SciPy library (Jones, Oliphant, and Peterson 2001) for statistics, and the matplotlib library (Hunter 2007) for plotting capabilities. The raw dataset, which was automatically extracted from the RateMyProfessors site, includes 1,231,643 entries with the following information: the instructor’s first and last name, department, institution, state, and city. Each entry also includes fields specifying the clarity, helpfulness, overall quality, easiness, and hotness scores, as well as the number of ratings. The discipline names within the raw dataset were modified to account for inconsistent naming conventions (e.g. ‘Physical Education’ and ‘Physical Ed’) throughout the RateMyProfessors database.

For the majority of this analysis (unless otherwise stated) only professors with a minimum of 20 ratings were considered, therefore reducing the raw dataset to a size of 7,882,980 ratings for 190,006 professors. This cut-off eliminates potential inaccuracies associated with professors having only a few submissions. All cut-offs used in this study were chosen such that any increase in their values did not change the numerical results by more than the amount of precision used to report the data.

For statistics related to hotness scores, any professor with a hotness score greater than zero was considered ‘hot’, whereas those with a hotness score of zero or below were considered ‘not hot’. Since the number of reviews varies greatly between professors, it is not meaningful to directly compare hotness scores. Weighting the hotness scores by the number of ratings is also not ideal, as the hotness score is based on an optional question in the submission form with no indication of how many students answered it. As such, in this study, the effects of attractiveness are based on whether a professor is rated as ‘hot’ (i.e. hotness score > 0) or ‘not hot’ (i.e. hotness score ≤ 0).

To predict professor gender based on first names as listed on RateMyProfessors, the R gender package was used (Mullen 2015). The R gender package compares a list of first names with historical data and produces a probability that the name refers to a male or female based on the reference dataset. In this study, the reference data was drawn from social security data by the U.S. Social Security Administration for individuals with birth years of 1930 and later (“Baby Names from Social Security Card Applications-National Level Data” 2015). Due to the potential gender ambiguity of some names, a cut-off probability of 99% was implemented; all professors with first names below this cut-off were disregarded for this portion of the analysis. This resulted in 73,004 male professors and 55,464 female professors with at least 20 ratings on RateMyProfessors. It should be noted that this gender-identification algorithm does have some limitations, as it may exclude certain demographics, such as professors with foreign names that are more likely to have less definitive and/or region-specific male or female usage (e.g. ‘Ming’, ‘Sasha’).

Results and discussion

Overview of dataset

To provide an overview of the dataset, box plots of the clarity, helpfulness, overall quality, and easiness scores for 190,006 professors on RateMyProfessors with at least
20 ratings are shown in Figure 1. The box plots illustrate that student-submitted reviews on RateMyProfessors tend to be more positive than negative, with the median clarity, helpfulness, and overall quality scores ranging between 3.80 and 3.86 on a scale from 1 to 5. This is in agreement with the conclusion of Bleske-Rechek and Michels (2010) that ratings on RateMyProfessors tend to be more positive than negative, which goes against a prevailing assumption that publicly available web-based student evaluations frequently carry a negative bias. The easiness scores, with a median value of 3.10 out of 5, do not have as much of a positive bias as the instruction quality ratings. Due to the larger sample size of this data set, the mean rating criteria as shown in Figure 1 differ slightly from the results of related studies (Otto, Sanford, and Ross 2008; Sonntag, Bassett, and Snyder 2009), although the overall positive bias in instruction quality scores is still in agreement.

![Figure 1. Box plots of the RateMyProfessors rating criteria for 190,006 professors with at least 20 ratings. The bottom and top boundaries of the box are the lower and upper quartile, respectively. The line within each box is the median, the square point is the mean, and the edges of the whiskers represent the lower and upper extrema.](image)

For histograms that visualize the relative frequency of scores for each rating criteria, refer to Figure S1 (clarity), Figure S2 (helpfulness), Figure S3 (overall quality), and Figure S4 (easiness) in the Supplemental Material. Figure S1, Figure S2, and Figure S3 highlight the left-skewed nature of the distribution of instruction quality scores, whereas Figure S4 shows that the easiness scores follow more of a normal distribution.

It is also worth investigating the overall quality scores of professors as a function of the number of ratings. An increased number of ratings could be due to aspects such as increased teaching experience, the popularity of the professor, the size of the class, or if the professor explicitly asks students to submit a rating on RateMyProfessors. Nonetheless, Figure S5 in the Supplemental Material shows that
there is no major trend in overall quality scores as a function of the number of ratings, indicating that professors with a higher number of ratings do not have any pronounced benefit or detriment compared to professors with fewer ratings. There is an approximate 0.1-point decrease in the average overall quality score between 1 and 20 ratings per professor, but this effect is minor and controlled for in this study since the data analysis only considers professors that have at least 20 ratings. For reference, the frequency that professors have a given number of ratings is displayed in Figure S6 in the Supplemental Material.

Since location-based data was also included in the dataset, the average overall quality scores of professors on a state-by-state basis was investigated. While some differences in average overall quality scores were present, no clearly discernible regional trends were observed. In addition, the number of professors with ratings on RateMyProfessors varies greatly between each state, with some states not having enough RateMyProfessors entries for statistically meaningful comparisons (e.g., Wyoming only has 8 professors with at least 20 reviews on RateMyProfessors).

**Correlations between rating criteria**

As a first test to observe the correlation between the various rating criteria, Pearson product-moment correlation coefficients and two-tailed p-values were computed, as shown in Table 1. The results of this test indicate that there is a statistically significant correlation between clarity, helpfulness, overall quality, and easiness scores on RateMyProfessors. The overall quality–helpfulness and overall quality–clarity relationships are very strongly correlated. This is to be expected, as overall quality is simply the average of both the clarity and helpfulness scores. The relationship between clarity and helpfulness is quite strong as well, indicating that students tend to be rather consistent when rating their professors on these two factors.

There is also a fair degree of positive correlation present between the clarity, helpfulness, and overall quality scores and the easiness score. These results are in agreement with the RateMyProfessors-related findings of Felton et al. (2008). The positive correlation between easiness and instruction quality from data on RateMyProfessors seems to go against prior research on traditional student evaluation scores, which has indicated that this correlation should be weak at best (McKeachie 1997; Marsh and Roche 1997; Marsh and Roche 2000). However, as noted by Otto, Sanford, and Ross (2008) this could be due to an ambiguous interpretation of easiness in the submission form. For instance, an instructor’s easiness may be interpreted as ‘easy to understand’ as opposed to ‘not challenging’.

<table>
<thead>
<tr>
<th>Clarity</th>
<th>Helpfulness</th>
<th>Overall quality</th>
<th>Easiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarity</td>
<td>—</td>
<td>0.94***</td>
<td>0.99***</td>
</tr>
<tr>
<td>Helpfulness</td>
<td>0.94***</td>
<td>—</td>
<td>0.98***</td>
</tr>
<tr>
<td>Overall quality</td>
<td>0.99***</td>
<td>0.98***</td>
<td>—</td>
</tr>
<tr>
<td>Easiness</td>
<td>0.60***</td>
<td>0.60***</td>
<td>0.61***</td>
</tr>
</tbody>
</table>

*** Correlation is significant at the p < 0.001 level (two-tailed).
Looking at these trends in more detail, the easiness rating as a function of clarity, helpfulness, and overall quality scores is shown in Figure 2. These values are from all professors in the dataset with at least 20 ratings. The data in this plot is divided into 0.1-point bins along the ‘average rating of instruction’ axis, and the mean (in both dimensions) is taken within each bin. This was done to reduce statistical noise and to address the discrete nature of the RateMyProfessors rating system. As anticipated from Table 1, there is a strong degree of positive correlation between a professor’s easiness and instruction quality. Of course, it should be noted that there is ambiguity regarding the causality, so it is unclear from this data alone whether students perceive better instructors as easier or if students perceive easier instructors as better at teaching. An additional trend from Figure 2 is that at the upper end of the instruction quality rating scale (i.e. above approximately 4.25), the slope in the average easiness score as a function of instruction quality is nearly twice as high as for professors with more moderate instruction quality scores.

Figure 2. Plot of average easiness scores among professors on RateMyProfessors as a function of their average rating of instruction, which includes overall quality (solid line), clarity (dashed line), and helpfulness (dot-dashed line).

Since overall quality is simply the average of clarity and helpfulness, it is interesting to consider why easiness scores correlate so strongly with these two factors. It is certainly possible that students consider a class to be of high quality if it is relatively easy, especially if the easiness of a course is a main deciding factor in the course selection process. Regardless of the underlying reason, RateMyProfessors inherently weighs the easiness score of a professor more-so than institutional student evaluations by not including other common measures of teaching effectiveness, such as the instructor’s knowledge and preparedness.
The fraction of professors rated as ‘hot’ as a function of average RateMyProfessors rating criteria is shown in Figure 3. Once again, the values are from all professors in the dataset with at least 20 ratings. The data are then divided into 0.1-point bins along the ‘average rating’ axis, and the mean (in both dimensions) is taken within each bin. The results of Figure 3 indicate that students tend not to perceive professors with poor instruction ratings as attractive, especially below clarity, helpfulness, and overall quality scores of 2.5 where the fraction of ‘hot’ professors is almost negligible. However, this fraction sharply increases with scores over 2.5 such that just over 70% of professors with perfect clarity, helpfulness, or overall quality scores are rated as ‘hot’ on RateMyProfessors. From low easiness scores up to a value of about 4.0, a nearly linear trend is seen in the fraction of ‘hot’ professors as a function of average easiness score. Above this point, the fraction of ‘hot’ professors is a relatively consistent 40%.

![Figure 3](image)

Figure 3. Plot of the fraction of professors rated as ‘hot’ as a function of average rating, which includes overall quality (solid line), clarity (dashed line), helpfulness (dot-dashed line), and easiness (dotted line).

Results by discipline

In this subsection, disciplines with a minimum of 200 professors that have at least 20 ratings were considered. The cut-off was selected as a balance between reducing the effect of entries with a small number of ratings while still having a sufficient sample size to accurately represent the major disciplines of study. After applying this cut-off to the dataset, the average overall quality scores for professors as a function of discipline were computed. The bottom and top 10 disciplines sorted by average overall quality score are shown in Figure 4. Notably, the majority of the lowest ranking disciplines in this regard are STEM fields. Contrastingly, all of the top 10 disciplines as sorted by overall quality are in the humanities, arts, or social sciences. Since there is such a strong
correlation between clarity, helpfulness, and overall quality, these trends are also true for the other metrics of instruction quality on RateMyProfessors.

Looking at easiness scores, it is clear that an analogous trend exists. Figure 5 displays the average easiness scores as a function of discipline for the bottom and top 10 disciplines. Once again, many of the most difficult disciplines as rated by students on RateMyProfessors are classified as STEM fields, whereas this is not true for the easiest disciplines on RateMyProfessors. As shown in Table 1 as well as Figure 2, there is a statistically significant dependence between overall quality and easiness, so it is not surprising that many of the disciplines included in Figure 4 are also included in Figure 5.

For a larger list of average clarity, helpfulness, overall quality, and easiness scores by discipline, refer to the formatted Microsoft Excel worksheet in the Supplemental Material. The dataset includes a total of 75 disciplines that have at least 100 professors with at least 20 ratings.
Results by gender

It is well-established that implicit gender biases play a role in how students perceive, interact with, and evaluate their instructors. Previous studies have found, for instance, that female instructors are perceived as friendlier but are judged more closely than male instructors (Bennett 1982; Costin, Greenough, and Menges 1971). In the context of RateMyProfessors, it has been found that the written comments on a professor’s page are strongly influenced by the gender of the professor, with positive words such as ‘brilliant’ and ‘genius’ appearing more often for male professors than female professors regardless of discipline (Schmidt 2015). Nonetheless, there is still much debate for whether this gender-stereotyping translates to student evaluations of teaching, and, if so, to what degree it impacts various criteria in student evaluations of teaching (MacNell, Driscoll, and Hunt 2014).

Using the R gender package (Mullen 2015) as outlined in the Methodology section, the average clarity, helpfulness, overall quality, and easiness scores of 73,004 male professors and 55,464 female professors were compared. To determine the statistical significance of the results, the outcomes for the male and female groups were compared using the Wilcoxon-Mann-Whitney two-sample rank-sum test. This is a nonparametric test of the null hypothesis that two samples come from the same population for two independent populations, in this case male and female professors on the RateMyProfessors website.

Figure 5. Easiness scores as a function of discipline for the bottom 10 (left) and top 10 (right) disciplines sorted by easiness for disciplines with a minimum of 200 professors that have at least 20 ratings.
The mean RateMyProfessors scores are summarized in Table 2, and all of the rating criteria have statistically significant differences between the male and female groups such that the null hypothesis could be rejected at the \( p < 0.001 \) level. On average, while male and female professors have statistically significant differences in ratings on RateMyProfessors, this difference in scores is relatively small. Female professors, on average, score 0.04 to 0.05 points lower on metrics of instruction quality and 0.03 points higher on easiness scores compared to male professors. These results appear to be consistent with prior systematic reviews (Feldman 1992; Feldman 1993) that found that gender differences in institutionally administered student evaluations of teaching were present but too small to be practically significant when considering average ratings.

<table>
<thead>
<tr>
<th>Rating criterion</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarity</td>
<td>3.71***</td>
<td>3.66***</td>
</tr>
<tr>
<td>Helpfulness</td>
<td>3.76***</td>
<td>3.72***</td>
</tr>
<tr>
<td>Overall quality</td>
<td>3.75***</td>
<td>3.69***</td>
</tr>
<tr>
<td>Easiness</td>
<td>3.10***</td>
<td>3.13***</td>
</tr>
</tbody>
</table>

*** The trend that one of the populations has larger values than the other is significant at the \( p < 0.001 \) level (two-tailed).

Nevertheless, it is essential to isolate known confounding variables in the dataset. As shown in Figure 4 and Figure 5, RateMyProfessors scores tend to be influenced by discipline. Since the gender ratio of faculty members varies significantly by field, this is a factor that needs to be accounted for. Another significant difference between the male and female faculty on RateMyProfessors is the fraction of professors in each subgroup that are rated as ‘hot’. Approximately 22.7% of male faculty on RateMyProfessors are rated as ‘hot’, whereas 27.8% of female faculty on RateMyProfessors are rated as ‘hot’. Since it has already been shown in Table 1 that perceived physical appearance correlates with evaluation scores, whether a professor is rated as ‘hot’ or ‘not hot’ should be controlled when analysing potential gender biases as well.

By only comparing ‘not hot’ (i.e. hotness score \( \leq 0 \)) professors by individual disciplines, the effect of gender biases on RateMyProfessors becomes much more apparent. As summarized in Table 3, each discipline has unique differences between ratings of male and female faculty members despite the small gender differences when looking at the RateMyProfessors dataset as a whole. Some disciplines, such as Chemistry, show no statistically significant difference in RateMyProfessors scores for male and female professors, whereas other disciplines, such as History and Political Science, have a large disparity. Importantly, out of all the popular disciplines on RateMyProfessors, there are no fields where women have statistically higher overall quality scores than men.
Table 3. Mean RateMyProfessors rating criteria for male and female faculty rated as ‘not hot’ in the top ten most popular disciplines on the site.

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Gender</th>
<th>Number of professors</th>
<th>Clarity</th>
<th>Helpfulness</th>
<th>Overall quality</th>
<th>Easiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Male</td>
<td>6287</td>
<td>3.64***</td>
<td>3.70***</td>
<td>3.67***</td>
<td>3.11***</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>7264</td>
<td>3.45***</td>
<td>3.52***</td>
<td>3.48***</td>
<td>3.02***</td>
</tr>
<tr>
<td>Mathematics</td>
<td>Male</td>
<td>8690</td>
<td>3.31*</td>
<td>3.46</td>
<td>3.39</td>
<td>2.93**</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>4831</td>
<td>3.35*</td>
<td>3.49</td>
<td>3.42</td>
<td>2.97**</td>
</tr>
<tr>
<td>Biology</td>
<td>Male</td>
<td>3948</td>
<td>3.46***</td>
<td>3.53***</td>
<td>3.50***</td>
<td>2.63***</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>3026</td>
<td>3.34***</td>
<td>3.42***</td>
<td>3.38***</td>
<td>2.55***</td>
</tr>
<tr>
<td>Psychology</td>
<td>Male</td>
<td>3170</td>
<td>3.69</td>
<td>3.65</td>
<td>3.67</td>
<td>3.16**</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2830</td>
<td>3.66</td>
<td>3.63</td>
<td>3.64</td>
<td>3.22**</td>
</tr>
<tr>
<td>History</td>
<td>Male</td>
<td>5018</td>
<td>3.72***</td>
<td>3.66***</td>
<td>3.69***</td>
<td>3.02***</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2098</td>
<td>3.46***</td>
<td>3.45***</td>
<td>3.46***</td>
<td>2.86***</td>
</tr>
<tr>
<td>Chemistry</td>
<td>Male</td>
<td>3268</td>
<td>3.26</td>
<td>3.40</td>
<td>3.33</td>
<td>2.58</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>1460</td>
<td>3.25</td>
<td>3.40</td>
<td>3.33</td>
<td>2.58</td>
</tr>
<tr>
<td>Communications</td>
<td>Male</td>
<td>2089</td>
<td>3.73***</td>
<td>3.73***</td>
<td>3.73***</td>
<td>3.29**</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>1819</td>
<td>3.51***</td>
<td>3.52***</td>
<td>3.51***</td>
<td>3.21**</td>
</tr>
<tr>
<td>Business</td>
<td>Male</td>
<td>2695</td>
<td>3.51***</td>
<td>3.53***</td>
<td>3.52***</td>
<td>3.08</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>1112</td>
<td>3.35***</td>
<td>3.37***</td>
<td>3.36***</td>
<td>3.06</td>
</tr>
<tr>
<td>Political Science</td>
<td>Male</td>
<td>2606</td>
<td>3.65***</td>
<td>3.62***</td>
<td>3.64***</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>862</td>
<td>3.46***</td>
<td>3.46***</td>
<td>3.46***</td>
<td>2.95</td>
</tr>
<tr>
<td>Economics</td>
<td>Male</td>
<td>2776</td>
<td>3.30***</td>
<td>3.37***</td>
<td>3.34***</td>
<td>2.84</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>761</td>
<td>3.18***</td>
<td>3.29***</td>
<td>3.24***</td>
<td>2.84</td>
</tr>
</tbody>
</table>

*** The trend that one of the populations has larger values than the other is significant at the \( p < 0.001 \) level (two-tailed).
** The trend that one of the populations has larger values than the other is significant at the \( p < 0.01 \) level (two-tailed).
* The trend that one of the populations has larger values than the other is significant at the \( p < 0.05 \) level (two-tailed).

It is possible that there are underlying variables not accounted for in the RateMyProfessors dataset that would explain this phenomenon on a discipline-by-discipline basis. Two such possibilities are the effects of reviewer gender, which is...
unknown due to the anonymity of RateMyProfessors, and the specific gender stereotypes associated with each discipline. In fact, studies that more carefully account for differences due to the interaction between professor gender, student gender, and discipline reveal a significantly more pronounced and complex effect of gender (Basow 1995) than studies that only looked at whether male and female faculty members receive different mean ratings. As such, given the limitations of the RateMyProfessors dataset, the cause of the gender-related differences shown in Table 3 is not immediately clear and requires further study, perhaps by comparing the results with estimated student gender ratios and student perceptions of gender differences within each discipline.

For a larger list of average clarity, helpfulness, overall quality, and easiness scores sorted by gender and discipline, refer to the formatted Microsoft Excel worksheet in the Supplemental Material. This data represents the average ratings for 75 of the most popular disciplines on RateMyProfessors. The dataset includes professors with at least 20 ratings, who are rated as ‘not hot’, and whose gender could be determined with 99% confidence. The Mann-Whitney U values along with effect size estimates and 95% confidence intervals based on the methods described by Newcombe (2012) are shown in Table S1 of the Supplemental Material.

Conclusions

The findings in this study shed light on the most popular online source of student evaluations of teaching, RateMyProfessors.com. By considering the evaluations of 190,006 professors on RateMyProfessors who teach in the United States that have at least 20 ratings, a number of trends were observed. Statistically significant correlations between clarity, helpfulness, overall quality, and easiness scores were observed (Figure 1 and Table 1). All three metrics of instruction quality (i.e. clarity, helpfulness, and overall quality) positively correlate with average easiness scores (Figure 2), and the fraction of professors rated as ‘hot’ is drastically higher for professors with high RateMyProfessors scores than those with poor ratings (Figure 3). When considering trends by discipline, professors in STEM and other technical disciplines receive both worse instruction quality ratings (Figure 4) and easiness scores (Figure 5) compared to disciplines in the humanities and arts. In cases where standardized student evaluation forms are adopted across a given institution, it may not be fair to directly compare professor rating across disciplines. For instance, the average overall quality score for professors in Chemistry is 3.47 but is 3.90 for professors in Psychology. A discipline’s perceived easiness may explain much of this variation, although it does not account for all of it. While the underlying reason for differences in ratings across disciplines cannot be definitively determined in this study, there are multiple possibilities. It may be that students in certain disciplines hold their professors to different standards, perhaps due in part to what they anticipate from their courses. It could also be that certain subject matters are inherently more difficult to teach or that instructors in a given discipline have different priorities and training with regards to teaching. As such, one should be hesitant to make direct comparisons between instructors across disciplines based on student evaluations of teaching alone. In the case of institution-wide teaching awards, for example, student evaluations of teaching should not be the only metric used to evaluate teaching effectiveness, as it is particularly prone to discipline-specific biases.
When considering the effect of professor gender on average rating criteria on RateMyProfessors, it initially appears that there are small but practically indistinguishable differences (Table 2). Nevertheless, when controlling for teaching discipline and rated attractiveness, the differences in scores for male and female professors vary significantly by field (Table 3) and is likely due to underlying variables not present in the RateMyProfessors dataset, such as reviewer gender and/or gender-typing of certain disciplines. It is difficult to discern the underlying cause of these gender differences in the various disciplines, as female professors have lower scores than men in some fields (e.g. History and Political Science) but have no statistically significant difference in scores in other fields (e.g. Chemistry). That being said, it still appears that women are at a particular disadvantage when it comes to student evaluations, as there are no disciplines where women have statistically higher overall quality scores than men. It is difficult to make a broad conclusion that applies to gender-specific differences across all disciplines, but these findings imply that the fair interpretation of student evaluations of teaching requires an acknowledgment of differences due to gender biases. Score-related gender differences for each discipline, similar to the data included in this work, can be used to adjust for these gender disparities in the interpretation of student evaluations as measures of teaching effectiveness.

The trends in the RateMyProfessors dataset that have been presented in this work likely exist, at least in part, in institutionally administered student evaluations of teaching as well. Identifying the correlations and potential biases that exist in widely used, publicly available web-based faculty rating websites such as RateMyProfessors will provide a greater understanding of student evaluations of teaching.

**Acknowledgments**

The author would like to gratefully acknowledge Scott Nicholson for his constructive feedback and assistance with the data collection process.

**References**


