Rate My Professors: A Study Of Bias and Inaccuracies In Anonymous Self-Reporting

Alexander Katrompas
Department of Computer Science
Texas State University
San Marcos, TX 78666
Email: amk181@txstate.edu

Vangelis Metsis
Department of Computer Science
Texas State University
San Marcos, TX 78666
Email: vmetsis@txstate.edu

***Abstract*—Commentary-driven and opinion-driven "ratings" websites such as Yelp, Amazon Reviews, and Rate My Professors are ubiquitous on the Web. These review sites engage in voluntary self-reporting which is well known to be fraught with bias and inaccurate representations. This study is an investigation into the website known as Rate My Professors which ostensibly seeks to collect and report on college and university professor quality.**

Rate My Professors (RMP), like many review sites, is anonymous. While it is commonly believed that anonymity increases accuracy, studies have shown otherwise. Studies using anonymous self-reporting are well known to be unreliable, so much so that researchers are required to take compensatory measures to validate their results. Rate My Professors takes no such measures, and in fact there is no guarantee that a "student reviewer" is even a student.

This study investigates Rate My Professors for bias, inaccuracy, and invalid data. The study will show compelling evidence which supports the idea that anonymous self-reporting, without compensatory validation measures, is flawed and unsuitable for use in a decision making process.

***Index Terms*—Reviews, Anonymous Reviews, Ratings, Online Ratings, Professor Evaluations, Rate My Professors, Self-Reporting, Review Bias, Self-Reporting Bias.**

I. Introduction

The rise in social media, and the commentary-driven interactions of social media, has had great impact on products and services of all types. These social media review sites by their nature engage in voluntary self-reporting, which is well known to be fraught with bias and inaccurate representations. Voluntary self-reporting is also known to be overwhelmingly populated by those who have strong feelings on the matter-at-hand, and typically strong negative feelings generate more voluntary self-reporting than positive feelings [1], [2], [3]. Therefore, while the general population may be satisfied in a product or service, a few very unsatisfied reviewers may dominate the discourse. In response to that, many products and services have engaged in self-generated positive reviews to "balance the scales," which causes further inaccuracies in reviews [4], [5].

In addition, review sites like Rate My Professors (RMP) are anonymous. It is commonly believed that anonymity increases accuracy but studies have shown otherwise. Complete anonymity decreases accountability, and encourages overly emotional responses, thereby decreasing motivation to answer thoughtfully and precisely [6]. Moreover, when the commentary is from a person in a perceived subordinate relationship, complete anonymity encourages the commenters to comment in a way that increases their perceived control regardless of the truth [1], [7]. In other words, anonymity encourages bias in the form of the "revenge post" when a commenter feels both out of control and slighted [7].

Studies using self-reporting are well known to be unreliable, so much so that researchers are required to take compensatory measures to validate their results or risk having their work discredited. [6], [8], [1], [3], [9]. Anonymity further calls into question a study's validity, requiring further compensatory measures [6]. Review sites such as Rate My Professors take no such measures, and in fact there is no guarantee that a student reviewer is even a student.

This study seeks to investigate Rate My Professors for bias, inaccuracies, and invalid data to either support or disprove the implications of anonymous self-reporting without compensatory validation. To accomplish this, data was obtained both from Rate My Professors and from official university semester student evaluations. Rate My Professors data is both analyzed against itself for bias, and against official university reviews which are taken to be the ground truth (GT).

Our analysis shows that Rate My Professor data is overwhelmingly more negative than ground truth, has far wider variability, and is unsuitable for use for professor evaluation. In addition, we find a strong negative correlation between the perceived difficulty of the course and the perceived quality of the professor. This calls into question whether students are simply reacting to difficulty by rating professors poorly to enhance their perceived control of the situation [7]. Finally, we find trends that indicate women in STEM fields are rated with a stronger correlation between difficulty and quality, possibly indicating bias against women in STEM.

The rest of the paper is organized as follows. In section II we discuss data collection and organization. In section III we generate and analyze various correlation metrics, across various categories of the data, investigating the hypothesis of bias and inaccuracy in anonymous self-reporting. Next we present section IV, further analysis of the ground truth data as it relates to and represents RMP data. Finally we state our conclusions in section V.

II. Data and Methods of Retrieval

A. Ground Truth Data

The standard in professor evaluation by students in the US is the official college semester evaluation process. All accredited colleges and universities are required to provide students with an opportunity to evaluate the course and the professor. These evaluations are conducted under strict guidelines: the professor may not be present during the evaluation, the professor may not handle the completed evaluations, and those completing the evaluations must be registered and present during the class being evaluated.

The reviews are anonymous; however, there is an acknowledged and significant psychological difference between the anonymity of official semester evaluations and the anonymity of ratings sites such as Rate My Professors. Despite being anonymous, the in-class college evaluations are taken more seriously, and are perceived as being official and therefore possibly non-anonymous. In other words, despite an assurance of anonymity, students behave as if the reviews are non-anonymous due to the official and in-person nature of the evaluations. This is an excellent example of a compensatory measure taken to reduce bias and encourage honesty and accuracy in self-reporting [10].

In addition, significant research and effort has been conducted in the construction of student evaluations to reduce bias, encourage honesty, and provide a valid data set for institutional analysis. In short, college and university student evaluations are constructed by researchers in higher education, psychology, and statistics with the explicit goal of providing valid, true, and usable data [11]. This is a far cry from the casual and for-profit nature of review sites like RMP.

As such, official college evaluations are considered to be truthful and accurate, when aggregated evaluations based on a sufficient number of students are used [10], [11].

For the purpose of this study, 250 random official university student evaluations of professors were acquired manually and compiled into a form so as to be comparable to Rate My Professors data [12]. The source of data was the University of South Florida (USF), which makes their evaluation data public. Each USF professor collected also has a corresponding RMP profile with at least 3 evaluations. For each professor collected, all data for all semesters was acquired and compiled to a single "quality" score, directly comparable to the corresponding "quality" score of the professor's RMP profile.

B. RMP Data

Rate My Professors data was collected directly from the Rate My Professors website using a web crawler (created by one of the authors of this study) named Prof-Bot (previously named ProfessorXBot) [12]. A working demonstration of this crawler can be seen on the site Academic Integrity Group[1]. Each profile is parsed and stored/updated in a central database and connected to the website Academic Integrity Group. At the time of this writing, 100% of the profiles (1.73M) on RMP have been downloaded, processed, and stored.

From the total of 1.73M profiles, 250,000 were extracted randomly for this study. This was further broken down into the categories male, female, hard-area, soft-area, male soft-area, male hard-area, female soft-area, female hard-area (see section II-D for descriptions of the categories). Tables I through III show the breakdown of the random sampling [12].

Gender	Count	Ratio
male	134,295	.64
female	115,705	.46
total	250,000	1.0

TABLE I
RMP data, male versus female.

Area	Count	Ratio
hard	85.332	.34
soft	164,668	.66
total	250,000	1.0

TABLE II
RMP data, hard-area versus soft-area.

Gender	Count	Ratio
male hard-area	51,680	.21
female hard-area	33.652	.13
male soft-area	82,615	.33
female soft-area	82,053	.33
total	250,000	1.0

TABLE III
RMP data, male versus female in hard-areas and soft-areas.

C. RMP Information Retrieval

Rate My Professors information retrieval and the parsing of the data was challenging. RMP uses (at the time of this writing) a complex dynamic Web application structure presumably designed to thwart Web crawlers. Therefore, we do not wish to disclose the algorithm developed to retrieve the data as a matter of proprietary information. However, we can assert Prof-Bot achieved a 99.9% success rate downloading information with 100% accuracy.

D. Data Enhancement

Both ground truth data and RMP data have been enhanced with gender information based on first name. This assignment is based on an analysis of birth records in the United States from 1909 to 2018. Using this data, a probability score was derived based on the total number of babies born per name, per gender. Any score above 92% is considered reliable to assign gender to a first name. Professors below 92% gender certainty were not used in the study.

Both ground truth data and RMP data have been enhanced with the labels "hard" (inherently quantitative) versus "soft" (inherently qualitative) disciplines. Hard disciplines are taken to be any STEM field (or related area) and any field for which quantifiable metrics are inherently used in the discipline. This varies slightly from the traditional definition of what is considered a "hard science." For example, while economics is not generally considered a "hard science," and is in fact referred to as a "soft science," it is included in this definition of "hard"

[1] academicintegritygroup.com

because economics by its nature is inherently quantitative and requires significant understanding of high-level mathematics. Similarly, accounting and finance are included in the "hard" category because they are inherently quantitative, despite not being natural sciences. Conversely, general business studies including fields such as marketing and management, which inherently include a qualitative component, were not included in the "hard" category. Using this definition, all "hard" disciplines were marked as such, and anything not marked as "hard" was marked as "soft."

Table IV shows a sample of the data collected and enhanced. In the columns "gender" and "discipline" a f/m represents female/male and a S/H represents soft/hard respectively. The rest of the columns represent discipline (area), RMP identifier (RMP ID), the number of RMP ratings (Ratings), the RMP quality rating (Quality), the RMP difficulty rating (Diff), the average of official college evaluations over all semesters taught (Avg), and the weighted average of official college evaluations over all semesters taught (wAvg).

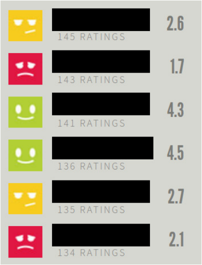
III. Correlation Analysis

A. *Ground Truth versus RMP*

Table V shows averages, standard deviations, and correlations comparing RMP data and official USF data for all categories. It can immediately be seen in the RMP data that average "quality" is considerably more negative across all categories (on average -10.01%), supporting the studies that anonymous, self-reporting promotes the propagation of strong negative feelings more than positive feelings [1], [2], [9]. It should be noted that the quality score averages cross the boundary from 3 to 4 in all cases. This is significant because that boundary is qualitative on RMP. When ranking professors, students are given qualitative measures like "average" (3) and "below average" (2). So the data clearly shows RMP data consistently on average ranks professors an entire qualitative category lower. This is especially egregious and misleading in that the RMP website presents color coding to show a professor is "average" or "below average" which impacts students' perceptions more than a purely quantitative presentation (see Figure 1).

The t-test in all cases shows there is significant difference in the means of the RMP data versus the ground truth data. This value is statistically zero in all cases, i.e. the Null Hypothesis does not hold. Therefore an alternative hypothesis needs to be made since the means appear to be significantly different and do not overlap. Collectively, initial observations seem to indicate that RMP data is a poor predictor and poor representative of the truth.

The next statistic of importance is the standard deviation. The variation in RMP data is considerably higher, on average 94.24% higher than the ground truth data. This is significant because it again supports the idea that anonymous, self-reporting is usually done by those with strong feelings and motivation creating greater variance. These "reports" are often done in the heat of the moment which can lead to high variance [1], [2], [9]. Official evaluations are done in a calm, academic setting, while RMP ratings are done when a presumed student is highly motivated to post a review.



🟨	[redacted]	145 RATINGS	2.6
🟥	[redacted]	143 RATINGS	1.7
🟩	[redacted]	141 RATINGS	4.3
🟩	[redacted]	136 RATINGS	4.5
🟨	[redacted]	135 RATINGS	2.7
🟥	[redacted]	134 RATINGS	2.1

Fig. 1. RMP quality color codes.

Correlation measures of the RMP and GT quality data are also interesting in a number of ways. First, the correlation between RMP and the ground truth in all cases is relatively weak, especially relative to the fact RMP also claims to be "truth." The highest is for males in "soft' areas at .75. The lowest is .58 for females in the "soft" areas (an interesting comparison in-and-of itself, worthy of further investigation). The average correlation for male and female in all areas is .65, which is relatively low. Taken alone, correlation is not low enough to say there is no relation, but combined with the strong difference in standard deviation, and the clear evidence that RMP quality data is significantly and consistently more negative, and that numerous studies show the clear flaws in anonymous self-reporting, we can take these relatively weak correlations as further evidence RMP data is likely flawed, biased, and inaccurate.

We also note that the highest difference in average ratings is for females in "hard" areas such as STEM fields, and the lowest difference is for females in "soft" areas such as art, languages, theater, etc. This is the first indicator that RMP is possibly facilitating bias against women in STEM fields. Indeed, for 17 years RMP had a "hotness rating" that was regularly used for harassment and encouraged the denigration of women, particularly in STEM fields[2] [3] [4].

B. *GT and RMP Quality versus RMP Difficulty*

The next statistic for analysis is the correlation between RMP's "difficulty" rating with the RMP quality and ground truth quality. The difficulty rating is highly subjective and is not defined by RMP. Presumably, it references how much "work" is required or how "difficult" the course is perceived. Quality is similarly not defined by RMP, however it is regarded by RMP as a student's rating of the "service" a professor provides as a teacher. Table VI shows these correlations for the GT professors.

It is notable that difficulty on RMP is negatively correlated with quality. In other words, students who rate a professor as "difficult" also tend to rate that professor as lower "quality." This is notable because it calls into question the sincerity of

[2] nypost.com/2018/07/10/professor-gets-ratemyprofessors-to-drop-its-hotness-rating/

[3] arstechnica.com/staff/2018/07/ratemyprofessors-drops-hotness-rating-prompts-me-to-rethink-undergrad-choices/

[4] www.buzzfeednews.com/article/juliareinstein/rate-my-professors-hotness-chili-pepper-sexist

Gender	Area	Discipline	RMP ID	Ratings	Qual	Diff	Avg
m	comm	S	2040473	31	4.8	2.2	4.6
m	math	H	1317726	161	2.6	3.7	3.73
f	f-lang	S	1551900	6	4.3	4.7	4.22
m	micro-bio	H	2233544	10	2.8	4.5	4.59
f	pub-health	S	1816109	4	4	3.5	4.23

TABLE IV
USF SAMPLE DATA WITH GENDER AND AREA ENHANCEMENTS.

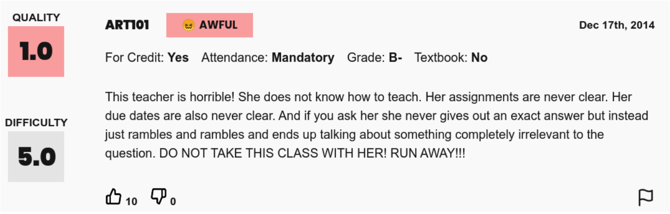
	All	Hard	Soft	Male	Female	M-Hard	F-Hard	M-Soft	F-Soft
RMP Mean	3.82	3.66	3.92	3.73	3.93	3.62	3.75	3.83	3.98
GT Mean	4.25	4.10	4.33	4.16	4.35	4.02	4.28	4.29	4.37
Difference	-10.01%	-10.78%	-9.60%	-10.34%	-9.65%	-10.00%	-12.41%	-10.62%	-8.80%
RMP Stdev	.94	.96	.92	.91	.96	.95	1.01	.88	.95
GT Stdev	.48	.49	.46	.51	.44	.46	.53	.52	.41
Difference	94.24%	96.49%	98.64%	79.78%	120.09%	107.84%	92.29%	68.16%	132.50%
Corr	.65	.62	.66	.68	.61	.60	.67	.75	.58
t-test	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

TABLE V
RMP VERSUS GT STATISTICS.

	All	Hard	Soft	Male	Female	Male-Hard	Female-Hard	Male-Soft	Male-Soft
RMP difficulty to RMP quality	-0.56	**-0.63**	-0.50	-0.58	-0.53	-0.57	**-0.74**	-0.58	-0.45
RMP difficulty to GT quality	-0.42	-0.40	-0.39	-0.45	-0.36	-0.31	-0.47	-0.49	-0.32

TABLE VI
RMP DIFFICULTY CORRELATED TO RMP QUALITY AND GT QUALITY. STRONGEST CORRELATIONS SHOWN IN BOLD HIGHLIGHTING A STRONGER RELATIONSHIP BETWEEN DIFFICULTY AND QUALITY FOR "HARD" AREAS AND FEMALE PROFESSORS.

the RMP quality rating as simply a proxy for how difficult the course was to complete (or vice versa). In other words, the question becomes: *Are students on RMP simply rating "difficult" courses as lower quality because they do not like difficult work?* The validity of this question is strengthened considerably when one notices the GT quality correlated to RMP's difficulty is significantly lower than RMP's quality correlated to RMP's difficulty. This could indicate that when a student rates a professor as low quality on RMP they are more significantly influenced by the difficulty of the course and are not honestly rating the quality of the professor (or vice versa). It may also indicate that students who rate a professor as low quality are purposefully rating the professor as high difficulty in order to increase their perceived control over the situation, and to steer other students away from a professor they may feel slighted them [7]. It is hypothesized that negative ratings on RMP typically infer, and often outright state, feelings such as "stay away from this professor!" It is reasonable to assume anyone making such a claim would be biased to also make the rating as deterring as possible to other students by rating the difficulty far higher than is warranted (see Figures 2 and 3).



QUALITY
1.0
DIFFICULTY
5.0

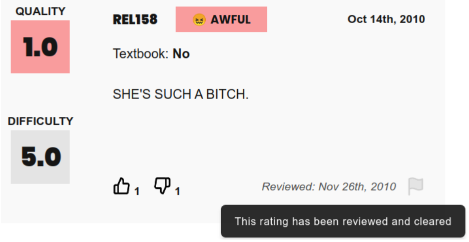
ART101 😖 AWFUL Dec 17th, 2014

For Credit: Yes Attendance: Mandatory Grade: B- Textbook: No

This teacher is horrible! She does not know how to teach. Her assignments are never clear. Her due dates are also never clear. And if you ask her she never gives out an exact answer but instead just rambles and rambles and ends up talking about something completely irrelevant to the question. DO NOT TAKE THIS CLASS WITH HER! RUN AWAY!!!

10 0

Fig. 2. Example comment showing low quality (1) with high difficulty (5) and the common ”stay away” message.



QUALITY
1.0
DIFFICULTY
5.0

REL158 😖 AWFUL Oct 14th, 2010

Textbook: No

SHE'S SUCH A BITCH.

1 1 Reviewed: Nov 26th, 2010

This rating has been reviewed and cleared

Fig. 3. Example comment showing low quality (1) with high difficulty (5), including a denigrating comment violating RMP's terms-of-service. Notice the comment is validated by RMP despite a clear violation of the terms-of-service, which is a common occurrence on RMP.

As further evidence of this relationship between quality and difficulty on RMP, there are two notable data points in that the two highest negative correlations are in "hard" areas, one of which is for both male and female (-.63), and the other is for female-only (-.74) (shown in bold). Also note the corresponding ground truth correlation is much lower in both cases. It is obvious there is a stronger negative relationship between the RMP difficulty and the RMP quality if the professor is in a quantitative field, particularly if the professor is female. Conversely, there is not as significant a relationship in the ground truth data, again calling into question the manner in which RMP gathers its ratings as being biased and inaccurate.

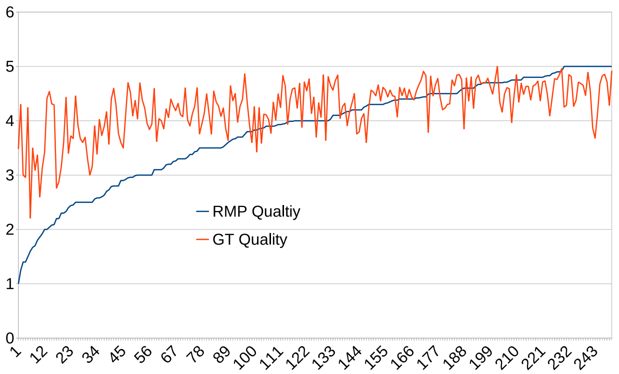


Fig. 4. Quality scores sorted on RMP quality (blue) with the corresponding GT quality (orange) for the 250 GT professors.

Admittedly, these correlations alone are not strong enough to draw general conclusions, but they add to the body of evidence of bias and inaccuracies, and certainly call into question the sincerity of the students on RMP. However, there are two correlation data points we can compare and draw stronger conclusions. The correlation of RMP difficulty and RMP quality for women in "hard" areas is significantly higher (-.74) than for any other category, and the same correlation in "soft" areas is significantly lower (-.45). This is notable because it indicates a bias on RMP against women in STEM while simultaneously demonstrating women in more "traditionally" female areas are treated more fairly. This is the second such indication of bias against women in this analysis, and this further supports the hostile, anti-female environment RMP has facilitated.

While these correlations are not strong enough yet to draw the definitive general conclusion that students on RMP are simply rating difficulty and not quality (we do not know the causal relationship), it certainly calls into question RMPs validity. Further, the aforementioned supposition that RMP reviews are biased against female professors is strengthened.

C. Graphical Analysis

In this section a graphical representation of the data is presented. Figure 4 shows the RMP and ground truth quality, sorted by RMP quality. It is clear that the ground truth quality is fairly evenly distributed, which is to be expected, while the RMP data is skewed wildly.

It is notable that the average quality on the ground truth data is slightly lower at the lower end of the RMP quality scores, implying some correlation to RMP, but there is not such a clear correlation at the higher end of the ratings. Observing this, correlations were calculated for the first, second, and third thirds of the sorted data. This resulted in the following:

- First third correlation (low ratings): .523
- Second third correlation (average ratings): .201
- Third third correlation (high ratings): -.018

These results suggest that while RMP quality rating has some weak similarity to ground truth data at lower quality rankings (1st third), professors ranked as lower quality are wildly misrepresented on RMP as far lower quality than the ground truth. Conversely, professors rated as high quality (3rd third) on RMP show no correlation at all.

This same analysis was performed on all the categories shown in Table VI with no significant difference in the results. Across all categories, all sorted results were almost identical. At the lower end of the RMP quality ratings, RMP ratings were extremely biased in the negative, and at the higher end of the RMP quality ratings, RMP ratings were generally in the same range but had almost no correlation other than being generally "high quality." We can conclude from this that the professors at the lower end of RMP quality ratings are wildly and unfairly misrepresented on RMP.

D. Validating GT Professors and Further RMP Analysis

In order to further analyze the connection between RMP quality and RMP difficulty, and to validate the connection between the relatively small sample size of the ground truth data and the larger RMP data set, two tables are presented. Table VII shows the correlations between RMP quality and RMP difficulty for various numbers of total ratings for the entire RMP data set. Table VIII shows the relationship between RMP quality and RMP difficulty for the GT professors on RMP as compared to all RMP data at various numbers of ratings. If the GT professors RMP data shows close relation to the overall RMP data set, we have confidence the comparisons presented here between RMP and GT are valid because GT professors are indeed a representative sample of RMP professors.

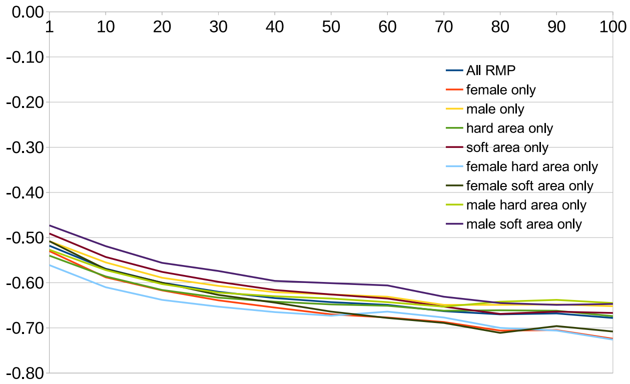


Fig. 5. Correlations of RMP difficulty versus RMP quality by Number of Ratings.

Table VII demonstrates that the negative relationship between RMP quality and RMP difficulty holds across all numbers of ratings. In fact, as the number of ratings goes up, the negative relationship gets stronger (see Figure 5). This implies that professors who generate higher motivation in students to rank their professor are also ranking with a stronger relationship between difficulty and low quality (and vice versa). This supports the hypothesis that voluntary self-reporting is highly biased and students are simply ranking difficulty as poor quality and/or high quality with ease.

ratings	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	over 100
rmpdata	-0.52	-0.57	-0.60	-0.62	-0.63	-0.64	-0.65	-0.66	-0.67	-0.67	-0.68
female	-0.53	-0.59	-0.62	-0.64	-0.66	-0.67	-0.68	-0.69	-0.71	-0.71	-0.72
male	-0.51	-0.56	-0.59	-0.61	-0.62	-0.63	-0.63	-0.65	-0.65	-0.65	-0.65
hard	-0.54	-0.59	-0.62	-0.63	-0.64	-0.65	-0.65	-0.66	-0.66	-0.66	-0.67
soft	-0.49	-0.54	-0.58	-0.60	-0.62	-0.63	-0.64	-0.65	-0.67	-0.66	-0.67
female hard	-0.56	-0.61	-0.64	-0.65	-0.67	-0.67	-0.66	-0.68	-0.70	-0.71	-0.73
female soft	-0.51	-0.57	-0.60	-0.63	-0.64	-0.66	-0.68	-0.69	-0.71	-0.70	-0.71
male hard	-0.53	-0.57	-0.60	-0.62	-0.63	-0.64	-0.64	-0.65	-0.64	-0.64	-0.65
male soft	-0.47	-0.52	-0.56	-0.57	-0.60	-0.60	-0.61	-0.63	-0.65	-0.65	-0.65

TABLE VII
CORRELATIONS OF RMP DIFFICULTY VERSUS RMP QUALITY BY NUMBER OF RATINGS.

	All	Hard	Soft	Male	Female	Male-Hard	Female-Hard	Male-Soft	Male-Soft
GT Professors on RMP	-0.56	-0.63	-0.50	-0.58	-0.53	-0.57	-0.74	-0.58	-0.45
All RMP Professors	-0.52	-0.54	-0.49	-0.51	-0.53	-0.53	-0.56	-0.47	-0.51
All RMP 1-10 Ratings	-0.57	-0.59	-0.54	-0.56	-0.59	-0.57	-0.61	-0.52	-0.57
All RMP 11-20 Ratings	-0.60	-0.62	-0.58	-0.59	-0.62	-0.60	-0.64	-0.56	-0.60
All RMP 100+ Ratings	-0.68	-0.67	-0.67	-0.65	-0.72	-0.65	-0.73	-0.65	-0.71

TABLE VIII
GT PROFESSORS AS COMPARED TO THE RMP DATA SET: DIFFICULTY VERSUS QUALITY.

Category	Half Category		One Category		Two Category	
Difference	Count	ratio	Count	Ratio	Count	Ratio
Negative	104	.416	51	.204	8	.032
Positive	15	.060	2	.008	0	.000
Total	119	.476	53	.212	8	.032

TABLE IX
GT PROFESSORS AS COMPARED TO THE RMP DATA SET: DIFFICULTY VERSUS QUALITY WITH RESPECT TO CATEGORY DIFFERENCES.

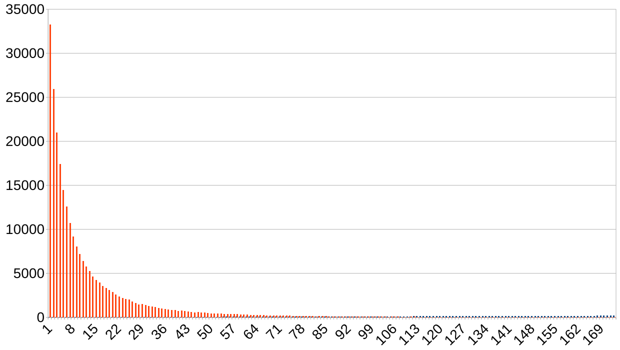


Fig. 6. Number of ratings per professor.

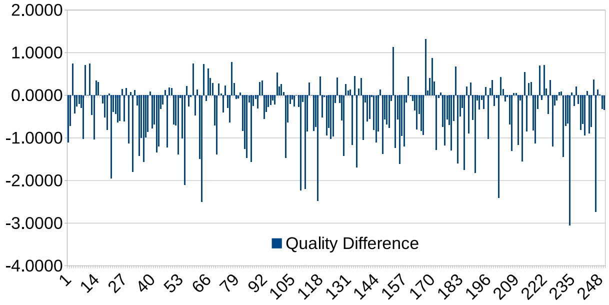


Fig. 7. Quality Difference RMP versus GT for the 250 GT professors.

Table VIII shows the RMP quality and RMP difficulty relationship for the GT professors on RMP, as compared to the entire RMP sample set at various numbers of ratings. The purpose of Table VIII is to demonstrate that although the GT professors on RMP are a relatively small sample size, they are in fact representative of the greater RMP data set. This table clearly shows that GT professors on RMP are representative of the greater RMP data set with very similar RMP difficulty to RMP quality relationships, with one notable exception; female professors appear not representative until the number of ratings posted is very high (over 90).

Further investigation into the numbers of ratings shows most professors have a relatively low number of ratings. Figure 6 shows the distribution of the number of ratings. By far, most professors have under 20 ratings. As can be seen in Table VIII the row "GT Professors on RMP" is comparable and similar with the rows showing ratings under 20 which is the common case. However, looking closer at number of ratings, we discovered female professors in hard areas are far more likely to generate ratings, and in many schools having an *average* number of ratings close to 70 with many over 90. Therefore, female professors in hard areas are also representative of the larger RMP data set when number of ratings are taken into account. It is notable that female professors are an outlier in both the strong negative relationship between RMP quality and RMP difficulty, as well as the number of reviews posted and the lower quality scores as compared to the ground truth.

We conclude that the GT sample set is representative of the RMP data set as a whole, and it is valid to then compare

GT quality data to RMP quality data. When this comparison is made, we can see that RMP quality is substantially lower than the GT, demonstrating RMP is both biased and inaccurate.

This analysis also points out the classical statistical bias present in most RMP data. Most professors have a very small number of ratings, which by definition is statistically invalid. Conversely, the ground truth data is an average across hundreds of ratings across many semesters which gives a more accurate view into a professor's true quality rating. The GT professors in this study have an average of 560 official ratings, which is therefore statistically valid and taken as truth.

Regardless of the causal relationship, it is clear that "difficulty" and "low quality" (and "easy" and "high quality") go hand in hand on RMP, and quality ratings in all cases are lower on RMP than in the ground truth. All analysis thus far supports the idea that anonymous self-reporting without compensatory validation is likely to be biased and inaccurate. It addition, there is strong evidence that female professors in quantitative disciplines are judged far more often and based far more on the difficulty of their field and not their teaching ability.

IV. Further Ground Truth Analysis

Figure 7 shows the differences in quality as reported on RMP versus the GT. It is immediately clear that RMP ratings are more extreme in positive and negative directions, but obviously far more so in the negative. Table IX shows the differences in this discrepancy between RMP and the ground truth. Almost half (47.6%) the professors are inaccurate by at least half a quality rating, and 87.4% of that difference is in the negative direction. A little over one fifth (21.2%) of the professors are rated an entire quality rating different, and 96.2% of that is in the negative direction. 3.2% of professors are rated 2 quality levels different from the ground truth and 100% are in the negative direction. An analysis of all the categories as shown in Table VI show comparable statistics for all categories, further demonstrating RMP quality ratings are significantly more negative than the ground truth.

V. Conclusions and Future Work

Rate My Professors states the following regarding their service: "*The site does what students have been doing forever, checking in with each other, their friends, their brothers, their sisters, their classmates, to figure out who's a great professor and who's one you might want to avoid.*" As such RMP is implicitly claiming their data is true and worthy of use in an important decision making process. They are claiming their reports are accurate enough to drive intelligent decision making in selecting a professor.

The combination of analysis presented here is consistent and clearly indicates RMP data is flawed, inaccurate, biased, and does not represent the truth. The flaws and bias found in RMP reporting is consistent with everything we know about anonymous self-reporting. Rate My Professors data is shown to be far more negative than the ground truth. This negativity is far stronger for professors at the lower end of the quality spectrum, giving an extreme and unfair bias to those professors. We found indications that gender is a factor and it is clear that gender bias exists on Rate My Professors. Further, this bias appears to be strongest for women in STEM. A specific study of female professors on RMP is warranted. In addition, although race was not studied in this analysis, this is also an area for future research. Both gender and race have been found to negatively influence students when rating women and minorities in official settings [13], [14]. Based on this study, we would expect this bias to be substantially worse in an anonymous self-reporting environment such as Rate My Professors.

Rate My Professors data is shown to be a poor approximation of the ground truth despite there being a weak correlation between the two. This is consistent with analysis that shows that despite a weak overall correlation, the larger variation and lack of consistency in RMP data makes it invalid for decision making.

Acknowledgment

The authors would like to thank Dr. Byron Gao for his guidance on data collection, storage, and parsing. In addition we would like to thank Jessica Katrompas and Julie Engleking for their extensive and tireless data collection efforts.

References

[1] S. Fulmer and J. Frijters, "A review of self-report and alternative approaches in the measurement of student motivation," *Educational Psychology Review*, vol. 21, pp. 219–246, 09 2009.

[2] P. Brenner and J. Delamater, "Lies, damned lies, and survey self-reports? identity as a cause of measurement bias," *Social Psychology Quarterly*, vol. 79, 11 2016.

[3] R. Rap and P. Paxton, "How accurate are self-reports of voluntary association memberships?" *Sociological Methods & Research*, p. 004912411879938, 10 2018.

[4] E. Anderson and D. Simester, "Reviews without a purchase: Low ratings, loyal customers, and deception," *Journal of Marketing Research*, vol. 51, pp. 249–269, 06 2014.

[5] J. Malbon, "Taking fake online consumer reviews seriously," *Journal of Consumer Policy*, vol. 36, 03 2013.

[6] Y. Lelkes, J. Krosnick, D. Marx, C. Judd, and B. Park, "Complete anonymity compromises the accuracy of self-reports," *Journal of Experimental Social Psychology*, vol. 48, pp. 1291–1299, 11 2012.

[7] N. Kowai-Bell, R. Guadagno, T. Chase, N. Preiss, and R. Hensley, "Rate my expectations: How online evaluations of professors impact students' perceived control," *Computers in Human Behavior*, vol. 27, pp. 1862–1867, 09 2011.

[8] K. Ritter, "E-valuating learning: Rate my professors and public rhetorics of pedagogy," *Rhetoric Review*, vol. 27, pp. 259–280, 07 2008.

[9] A. Stone, *The science of self-report. Implications for research and practice*. Lawrence Erlbaum Associates Publishers, 01 2000.

[10] D. Feistauer and T. Richter, "How reliable are students' evaluations of teaching quality? a variance components approach," *Assessment & Evaluation in Higher Education*, pp. 1–17, 11 2016.

[11] H. Marsh and L. Roche, "Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility," *American Psychologist - AMER PSYCHOL*, vol. 52, 11 1997.

[12] B. Gao and A. Katrompas, "A preliminary experimental analysis on ratemyprofessors," 12 2020.

[13] S. Basow, S. Codos, and J. Martin, "The effects of professors' race and gender on student evaluations and performance," *College student journal*, vol. 47, pp. 352–363, 06 2013.

[14] A. Bavishi, J. Madera, and M. Hebl, "The effect of professor ethnicity and gender on student evaluations: Judged before met," *Journal of Diversity in Higher Education*, vol. 3, pp. 245–256, 12 2010.