Journal of the Scholarship of Teaching and Learning, Vol. 24, No. 4, December 2024, pp.1-21.
doi: 10.14434/josotl.v24i4.33932

For Good Reason: Analyzing How Students Define Difficulty in RateMyProfessor.com Comments

Alexis Teagarden
University of Massachusetts Dartmouth
ateagarden@umassd.edu

Michael Carlozzi
University of Rhode Island
michaeljay@uri.edu

Abstract: Can student comments help solve a problem that student ratings helped create? We argue that the comment section of student ratings of instruction (SRI) offers a rich site for studying student perspectives on teaching and learning, particularly how students define and value course and instructor difficulty. Employing rhetorically grounded approaches to computer-assisted corpus analysis, we compared 4,600 RateMyProfessor.com instructor profiles meeting the criteria of 1) instructors with high difficulty and high overall quality scores or 2) instructors with high difficulty but low overall quality scores. We identify recurring argumentative patterns in both corpora. In contrast to SRI scholarship which often assumes students favor ease over all other course characteristics, we see commenters providing a more nuanced evaluation: condemning artificial forms of difficulty but commending authentic ones. Our findings contribute to discussions of student perspectives on learning and their relationship to course evaluations. While we note SRIs should never be the sole means of evaluating faculty, the results offer evidence in support of the validity hypothesis in SRI scholarship and provide avenues for helping faculty better understand their course evaluations and students.

Keywords: student ratings of instruction, student evaluations of teaching, student perceptions, computer-assisted corpus-analysis

What do we mean by a "difficult course"? Readers of this journal could likely imagine examples. Perhaps they can also call to mind someone they would describe as "a difficult person." But what makes someone a "difficult professor"? Is the definition more related to the aspects of one's course or the aspects of one's personality? We suspect answers would vary, and even more so if we were to ask students.

Since learning depends on tackling increasingly challenging tasks or problems, understanding student perspectives on difficulty strikes us as vital. If instructors and students share definitions and values, difficulty can be a source of motivation. But, if views differ, then difficulty can lead to miscommunication and missed learning opportunities. And we cannot presume that faculty definitions will align with those of students; Lauer (2012), for instance, found that faculty and students differed in how they interpreted key words like "professional," "fair," and "respectful" on course evaluation forms.

Faculty evaluation forms, often called student ratings of instruction (SRI) or student evaluations of teaching (SET), have prompted many arguments about students' views of difficulty. Prolific SRI researcher Dennis Clayson (2014) suggested that "if students like an instructor (for whatever reason), then the easiness of the class becomes relatively irrelevant" (p. 695). Clayson et al. (2006) have even argued that SRIs "could be replaced with a personality inventory of the instructor with little change in outcome" (p. 158). For this line of argument, course difficulty or ease is presumed to be irrelevant to student judgment, and Clayson's almost fatalistic position suggests that instructors have little agency over their SRI ratings.

This represents an extreme denial of SRI validity, but beliefs about students' perceptions of difficulty form the basis of a core debate within the literature. SRI opposition is often premised on the claim that these data are corrupted by a student preference for easy courses, which is translated as a preference for learning less (e.g., Clayson et al., 2006; Stroebe, 2020). This "leniency hypothesis" tends to assume that the difficulty for which students "punish" instructors via low SRI scores is an educationally desirable trait. This perspective assumes that students favor ease over most if not all other course traits; it also suggests researchers understand what students mean by easy and difficult.

If student views of difficulty possibly shape their approach to learning, and certainly affect arguments about SRI, then we need grounded articulations–not assumptions–of student definitions. And since SRIs often elicit student views of difficulty, they provide an excellent resource.

Accessing campus-based SRIs remains challenging, however, so much work on ratings of faculty has turned to RateMyProfessor.com (RMP) data. Shifting from campus-based SRIs to RMP can appear a sleight-of-hand; in-class course evaluations differ from online ratings in multiple ways. Scholarship on RMP has addressed such concerns, documenting a substantive correlation between in-class and online scores (see, for example, Bleske-Rechek & Michels, 2010; Coladarci & Kornfield, 2019; Timmerman, 2008) as well as a correlation between grades and scores, which is also well-documented in campus-based SRIs (Carlozzi, 2018; Constand et al., 2016; Otto et al., 2008). RMP's massive collection of quantitative and qualitative data further allows for the creation of a corpus big enough to overcome issues like occasional ratings from as-students. In short, RMP has proved a useful stand-in for hard-to-access institutional SRIs.

RMP provides one further affordance; it prompts students to rate both overall quality and difficulty. The current interface's first and second question read, respectively: "Rate your professor" and "How difficult was this professor?" RMP research shows the site directly asked about difficulty/ease as far back as 2008 and kept this rating distinct from the quality score, even when quality was computed by averaging ratings from other questions (Otto et al., 2008). RMP also provides guidance on how to score difficulty. At the time of this writing, the site's help page explains the difficulty rating as "Some students want to know how easy or difficult a class is before they register. Is this class an easy A? How much work needs to be done in order to get a good grade?" This explanation appears similar to earlier versions (Otto et al., 2008). How well these guiding questions shape RMP users' actual scoring practices is another matter of assumption, however.

But we need neither to rely on assumptions nor content ourselves with silent quantitative scores. RMP's open-ended comments offer a way to check assumptions against actual statements by raters. Carlozzi (2018) suggested the potential for RMP comments to elucidate actual meanings of difficulty. While his analysis focused on instructor RMP scores in relationship to their SRI research agendas, Carlozzi noted:

> When rating difficult instructors on RMP, students focused on several characteristics that reflect ineffectual – and potentially absent – teaching. A class was rated difficult because, for example, the instructor refused to answer emails and to meet with students, referring all questions to teaching assistants. I argue that this does not reflect academic rigour. It portrays a difficult class, but one which is logistically, even artificially, hard. Further analysis of RMP comments would be needed, but the odds ratios in this paper, alongside my admittedly cursory review of RMP comments, suggest that classes are rated 'difficult' based in large part on the instructors' uninterested attitude towards students. (p. 7)

Carlozzi's findings were preliminary and speculative but suggested a meaningful path for understanding how students define difficulty. Our study explored how well Carlozzi's preliminary finding held, using RMP's focus on difficulty to ask how commenters justified their scores of overall quality and overall difficulty.

To develop our analysis we compared RMP comments left on the profiles of 4,600 total instructors, evenly split between two profile types: those rated high in difficulty and high in overall quality, and those rated high in difficulty but low in overall quality. To see how commenters explained RMP evaluations, we employ a mixed-methods approach by drawing on computer-assisted emotion and corpus analysis, investigating how the corpora differ generally and then with regards to traits of effective teaching.

Our analysis supports Carlozzi's hypothesis about rater definitions of difficulty. First, we found that comments often reveal what RMP raters mean by difficulty and, second, that commenters tacitly define two kinds of difficulty:

- Authentic difficulty includes work that students see as hard but also relevant to course outcomes and post-class goals: challenging assessments given within a supportive or inspiring learning environment, demanding but fair grading practices, and other combinations of difficult learning tasks paired with positive teaching traits (e.g., respectful, engaging, knowledgeable, communicative).
- Artificial difficulty includes logistical problems unrelated to course subject or design like repeated difficulties accessing class lectures/materials or perceived mismatches between what is covered in class and what is assessed in exams and assignments; such issues are also often tied to a perceived lack of positive teaching traits.

This differentiation represents a more nuanced understanding of difficulty than is assumed in much SRI scholarship. We see RMP commenters, while at times unable to articulate themselves well, able to make pedagogically reasonable evaluations of class difficulty. They commend instructors who present authentic difficulty and condemn those whose courses feature artificial difficulty.

Our findings also illustrate how attention to SRI commenting patterns can provide better ways of navigating these data in formative settings and summative evaluations. Showing faculty how students discern purposefully challenging assignments from work which is merely a logistical headache might instill the confidence to include harder material—or the motivation to remove artificial hurdles. Recognizing argumentative patterns within student comments can thus improve our teaching and the way we view SRIs, which we have argued can be a useful, if limited, element of a multi-faceted feedback system (Teagarden & Carlozzi, 2020).

In making this argument, we first summarize the relevant literature on difficulty in SRI and RMP research. We then describe and present findings from three analyses:

1. Emotion analysis
2. General keyword & corresponding cluster analysis
3. Teaching traits' keyword analysis

We conclude by discussing the presence of race and gender indicators in each of the corpora and offer recommendations about managing course difficulty.

Background: The Concept of Difficulty in SRI Scholarship

Student beliefs about difficulty underpin many arguments about SRIs. Perhaps the most visible example is the leniency-validity debate, which represents a foundational, if often unacknowledged, divide. Its stakes might be as high as the continued use of SRIs for summative and even formative teaching evaluation. This is because the leniency hypothesis discredits SRIs by holding that enough students engage in a form of quid-pro-quo with their faculty, rewarding (unmerited) good grades with (unmerited) good scores (e.g., Clayson et al., 2006; Stroebe, 2020). Leniency proponents base their argument on the repeated finding that students' grades correlate with SRI scores (Benton & Ryalls, 2016; Spooren et al., 2013; Wang & Williamson, 2022). If the leniency hypothesis argument is correct, then SRIs not only fail to measure good teaching, they also punish faculty who offer the rigorous courses we desire in college teaching, contributing to problems like grade inflation and, more generally, eroding our educational mission.

The leniency hypothesis faces opposition from the validity perspective. Validity advocates also grant the grade-scores correlation but offer a competing interpretation. Students with higher grades, they argue, likely learned more, and in turn, these students provided high ratings to the instructors that fostered this learning (Spooren et al., 2013; Wang & Williamson, 2022). From this perspective, SRIs measure exactly what they should and thus serve as valid instruments for improving teaching and, when used appropriately, evaluating instructors. The two camps have yet to resolve their dispute, perhaps because few articles consider how such debates hinge on an ill-defined concept of difficulty.

Research on SRIs has shown, however, that students like to be challenged. Benton et al. (2013), for example, find that "Students who perceive the instructor expects them to share in the responsibility for learning and sets high achievement standards are more likely to make progress in the course and assign high ratings" (p. 12). Marsh (2001) approaches the subject from another angle, stressing "the substantive importance of distinguishing between Good and Bad Workloads" (p. 206). "Bad" workloads are "imposed" without consideration of students' abilities or prior learning and with an unrealistic pace of delivery. "Good" workloads challenge students and are "substantially positively correlated with the Overall Teacher Rating factor [on SRIs]" (p. 197). To improve student ratings on SRIs, Marsh recommends improving "good" instructional hours, that is, course-related hours which students regard as valuable.

Marsh's findings on workload align with more general studies of traits associated with good teaching., which further suggests that students might have good reasons behind their ratings of difficulty. Hoyt and Lee (2002), for example, found that SRI scores correlated with certain types of teaching methods: showing an interest in students, helping to make subject matter relatable, introducing stimulating ideas, and delivering clear feedback and criticisms. Since these teaching methods are widely understood as facets of good teaching, Hoyt and Lee suggested that student ratings align with expert views on good teaching. Such work has not persuaded members of the leniency hypothesis perspective, perhaps because it is unclear how, or even if, students associate these general teaching traits with SRI questions.

Research on the Potential of SRI Comments to Clarify Student Definitions of Difficulty

Yet SRIs provide a means of analyzing how students define and evaluate difficulty in classes, since most campus-based SRIs and all RMP posts elicit open-ended comments. Such comments can elucidate rationales behind scoring decisions, including reasoning around ratings of difficulty. Prior research, furthermore, has suggested such comments merit attention. In one of the few studies to rigorously examine student comments, Brockx et al. (2012) revealed that comments reliably predicted SRI scores and mostly concerned a gap between theory and practice. Further, negative comments focused on "the evaluation and context categories," as opposed to positive comments, which focused on a course's various aspects, including the teacher (p. 1131). They concluded that students took commenting seriously.

The value of comments has also been shown in studies of RMP. Ritter's (2008) rhetorical analysis found that comments on five selected faculty pages "visibly discount the notion that only lazy or disgruntled students populate RMP" (p. 274). The posts she studied also include arguments over whether a professor was "hard" (p. 274). Ritter's work was challenged by Chaney (2011), who claimed that RMP, alongside larger social forces, leads students to value "easily consumable" education (p. 199). Chaney supported her claims by analyzing 100 literature and first-year writing faculty profiles. On the topic of difficulty, Chaney provided excerpts to show students praising instructors for easy classes and contrasted them with complaints about disorganized faculty, arguing the pattern reveals "an extreme impatience for any evidence of the professor's humanity that might be still infecting the classroom" (p. 199).

Extending this language-focused comment analysis, in a computer-assisted corpus analysis of Asian and non-Asian sounding RMP instructor profiles, Subtirelu (2015) found that commenters tended to reinforce dominant nativist ideologies relating to language but that commenters were also able to offer nuance through usage of an x+but construction, as in the phrase "he does have an accent but." This repeated construction illustrates one way that RMP commenters justify quality and difficulty scores for their perceived audience, and it presages one of our primary findings. While Ritter, Chaney, and Subtirelu come to differing conclusions, their close attention to the language of RMP comments suggests that RMP participants attempt to explain difficulty ratings, which might provide insight into what students mean when they call a course or instructor hard.

While such studies explore what students value in general, they do not define difficulty or explain its place among good teaching profiles. Thus, what difficulty means to students remains mostly a matter of assumption. Our analysis addresses this by analyzing how raters themselves justify difficulty and overall quality scores.

Corpora Design

To analyze how, if at all, commenters perceived instructor difficulty and its relationship to overall teaching quality, we built two corpora of RMP comments:

1. The low rated corpus (LRC): those who were rated as very difficult (4 or greater) but very poor in overall quality (2 or lower overall)
2. The high rated corpus (HRC): those who were rated very difficult (4 or greater) but excellent in quality (4 or greater overall)

These scores are cumulative. We also built a corpus of instructors sampled without qualifying criteria as a reference point.

The data collected required some cleaning; for example, we edited the texts to work better with the corpus software by removing the empty phrase "No comments," which RMP uses to indicate that a student left no comments. Other changes were made to remove characters that confused the software; for example, the software interpreted the "t" in "doesn't" as a separate word. We also replaced many "not" phrases with synonyms, e.g., "not fair" became "unfair" and "not boring" became "interesting." While these alterations might affect overall tone and reduce analytical precision, we aimed to identify patterns of reasoning rather than specific diction. By removing "not +" constructions, we were better equipped to trace common justifications; for instance, a student praising an instructor for being "not boring" and another student claiming that the instructor is "interesting" offer similar kinds of reasoning in our analysis, even if the comments differ rhetorically. We decided that the clarity of this change outweighed the small chance of miscategorization (e.g., cases where "not helpful" actually did mean "helpful"). The appendix contains the full list of textual changes.

Despite having the same number of instructors, the corpora had dramatically different word counts, suggesting that commenters had more to say about instructors whom they disliked: the edited LRC was 3,620,572 words and the edited HRC was 1,674,055 words. The average comment left on the LRC contained 49 words compared to the HRC's 40, and more students generally commented on the LRC (32 to 18 comments on average). This conflicted with previous research by Brockx et al. (2012) which had found that students left more positive than negative comments. They examined institutional SRI, however, and not RMP comments, and we did not sample instructors randomly, instead deliberately choosing a skewed rating distribution.

A Note About the Limitations of the Corpora Demographics

We caution against drawing inferences from our corpora descriptions. This caution governs our paper, from the overall approach to specific word choices. For example, given the size of our corpora, almost any difference noted among them will register as statistically significant. We doubt such differences have real-world meaning, nor does our corpora, as designed, account for well-established factors affecting SRI scores like academic discipline, course level, or required vs. elective status (Marsh, 2007; Benton & Ryalls, 2016), which would need to be accounted for before conducting inferential statistical analyses.

We did not aim to assess the effects of demographic variables on scores; instead, we checked only to see if there were differences in terms of order of magnitude. Our textual analyses also focused on substantial differences, ones that are obviously more than the product of sampling chance. We therefore intentionally avoid claims about statistically significant differences between corpora. We welcome researchers to run other kinds of analyses on our data; our corpora as well as the code used to generate and clean them are publicly available at [reference to journal-appropriate online repository].

Corpora Demographics

We used the Python package gender-guesser to classify the gender of instructors in our samples based on their first names. Such algorithms are prone to misclassifications and non-classifications, but gender-guesser performs well, having a dataset of over 45,000 names subject to review by international auditors (Santamaria and Mihaljević, 2018). We were interested primarily in locating *order of magnitude* differences between the HRC and LRC, knowing that these data are probably very noisy. In this sense, we cared more about misclassifications (assigning the wrong gender to a name) than non-classification (not being able to identify a name's gender). Gender-guesser suits this purpose; in a benchmark test, gender-guesser outperformed even commercial packages in terms of misclassification with an error rate under 3% (Santamaria and Mihaljević, 2018).

Gender-guesser assigns five states to a name: female, mostly female, male, mostly male, androgynous, as well as unknown (non-classification). For simplicity, and because they represented only 6% of the total classifications, we pooled the mostly female and mostly male classifications in their respective groups and then discarded the unknown and androgynous classifications.

We do not find meaningful differences between the corpora. The HRC did feature a larger percentage of male-identified instructors than the LRC: among those names able to be classified, 34% of instructors in the LRC registered as female compared to 31% in the HRC. We were also interested in how gender distributions differed from a general sample of RMP instructors, so we sampled 2,300 instructors on RMP without inclusion criteria and found females represented 30% of those members. We can conclude that female-identified instructors were slightly more likely to be found in the LRC compared to both the HRC and a reference group. We further discuss these findings, and the below demographic details, in our concluding section.

We next considered the race/ethnicity representation within the corpora, a much more difficult problem to tackle than gender, given the many ways in which race, culture, and ethnicity intertwine. For this classification, we used the Python package ethnicolr, which predicts race and ethnicity by names. We chose this package because its underlying dataset was the U.S. Census, a fairly credible repository for connecting names to races.

Ethnicolr is a probabilistic model and thus lists the probability (expressed as a percentage from 0-1) of a name belonging to a race and ethnicity. For example, it believes the name Jian Tang has a 98% chance of being classified as Asian/Pacific Islander (API) and the name Gerald Dougherty has a 95% chance of being classified as Non-Hispanic White. Vini Angel, though, is less clear, and the classifier believes it has a 27% chance of being Hispanic and 67% chance of being Non-Hispanic White. The classifier also assigns a best guess for the person's race, defined as the race with the majority of probability (usually but not always 50%+) assigned to it, but this statistic does not capture the uncertainty in a guess and so we report instead the mean and median probability of a name belonging to the different ethnic categories.

The probability of an instructor's name being API is 7% in the LRC but 4% in the HRC. The probabilities of Black names were identical in both corpora (7.7%), and Hispanic names were virtually identical (LRC = 5%, HRC = 4.4%). Median scores were basically indistinguishable in both corpora for all groups.

We also reviewed subject discipline representation. RMP classifies each instructor as being part of a discipline. We combined disciplines that had very small representation and were obviously part of a larger discipline (e.g., microbiology to biology); all of these changes are in the appendix, and they were performed on disciplines which represented less than 1% of the total dataset.

The LRC is dominated by mathematics, whose instructors alone represent 18% of the dataset; the next closest disciplines are computer science (8%) and English (8%). The distribution of disciplines in the HRC is far more even, with English first (10%), followed by biology (8%) and fine arts (7%). The sample of instructors without inclusion criteria is distributed more evenly, with subjects appearing to represent common and required courses; English is first (9%), followed by mathematics (9%) and science (7%).

All RMP comments included below are reproduced in their original form, including typos.

Analyses

Emotion Analysis

We first compared the overall feelings expressed within each corpus through emotion analysis. This analysis attempts to identify the commenters' collective emotional experiences, in contrast with sentiment analysis, which merely reports on valence, i.e., positive vs. negative language (Kušen et. al, 2017). Given the disparity in overall scores between the corpora, we did not believe sentiment analysis would be particularly helpful (we should expect the HRC to be overwhelmingly positive); emotion analysis, though, might help identify which emotions best captured the writing's mood. The appendix contains all R code used for these analyses as well as methodological details.

The emotions expressed by students differed widely according to corpus (Figures 1 and 2). Almost half (47%) of the HRC's emotions reflected joy or trust; anger, disgust, fear, and sadness totaled 31%. Even after we had removed many false positives, "trust" highlighted the HRC's emotional map. In contrast, negative emotions dominated the LRC (62%), with just 23% of emotions expressing joy and trust.

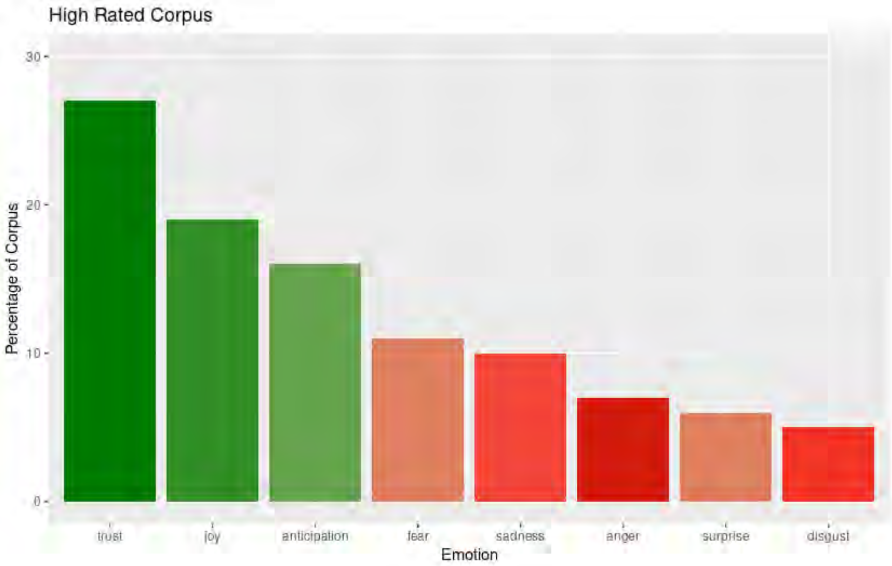

Figure 1. Emotions in the HRC as expressed in the NRC Emotion Lexicon as a percentage of the total corpus.

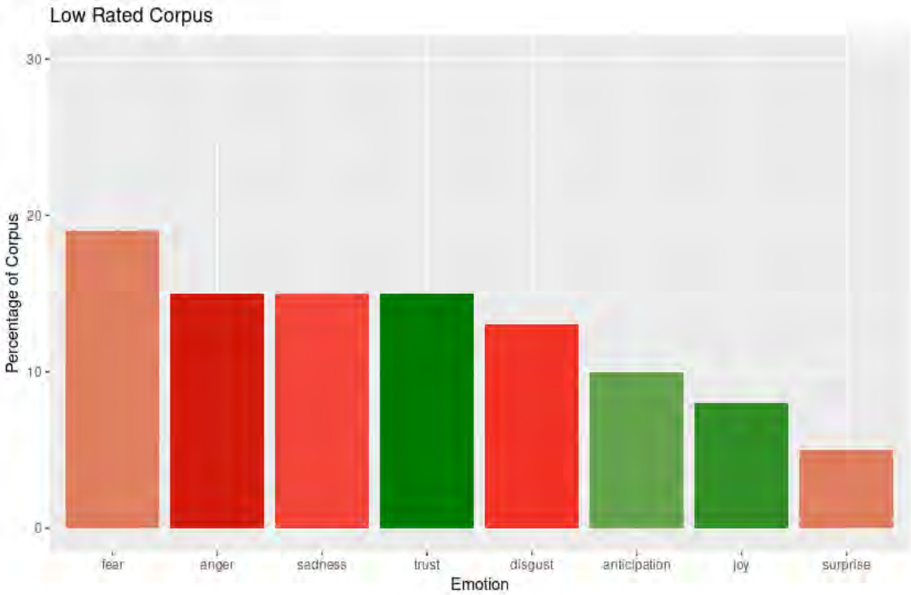

Figure 2. Emotions in the LRC as expressed in the NRC Emotion lexicon as a percentage of the total corpus.

These emotion analyses show that comments are aligned with their overall quality rating: comments in the HRC tended to express positive emotions whereas comments in the LRC expressed negative emotions. This particular opposition foregrounds challenges raised to the leniency hypothesis. Advocates of the leniency hypothesis assume that students base their ratings on grading ease, and so harder classes receive lower SRI marks. Presumably, class rigor would correspond to comments full of negative emotions. The LRC conforms to this prediction, with fear as the dominant emotion. Yet the HRC's instructors engender positive reactions, especially trust. How do those instructors create a positive relationship in spite of the commenter-acknowledged course difficulty? We turned to keyword analysis to examine whether comments revealed clear patterns of explanation.

General Keyword & Corresponding Cluster Analysis

Keyword analysis compares the occurrences of words between a "reference" corpus and a "study" corpus to determine the significance of observed differences in frequency. Often the reference corpus is a larger language set from which the study corpus comes, but that is not always the case, and in our analysis the reference corpus and study corpus were simply the HRC and LRC. Given this design, the frequency of words in one corpus stands out in relation to another corpus of instructors that students had also rated as highly difficult.

We used the freeware concordance software, AntConc, for keyword analysis (Anthony, 2020). AntConc allows for a deeper contextual understanding of word usage through clusters and collocates, as opposed to a basic frequency analysis. We used the text editing program Atom to determine word and cluster counts. When discussing frequency, we normalized to usages per million due to differing corpus sizes, and we denote this usage by /m.

Keyword analysis shows that commenters highly approved of the HRC instructors and highly disapproved of the LRC—predictably, given the quality scores. Perhaps also predictably, the top keywords for the LRC include "worst," "horrible," "terrible," and "bad," while the HRC's top keywords are "best" and "great" (Table 1).

Table 1. Comparison of top ten keywords in each corpus. Bolded findings discussed below.

LCR Top Ten Keywords	HRC Top Ten Keywords
1. worst 2. not 3. does 4. horrible 5. avoid 6. **teach** 7. unclear 8. terrible 9. bad 10. this	1. best 2. great 3. amazing 4. **but** 5. dr 6. awesome 7. interesting 8. professor 9. lot 10. helpful

The top 100 keywords also clarify that the HRC's commenters intended their high difficulty ratings, as they include seven related terms: tough, challenging, hard, demanding, hardest, difficulty, and challenges. The LRC's top 100 includes only "impossible" (analyzed below). Counter to the more extreme views of the leniency hypothesis, RMP commenters highly praised classes that they also describe as difficult. The key way LRC commenters discuss difficulty, on the other hand, meshes with the salient background emotion of fear.

Argument Patterns in the HRC: "difficult/synonym + but"

The HRC's basic keyword list shows frequent comments on difficulty. The fourth ranked keyword "but" points, however, to an argument pattern. Analyzing clusters shows the most frequent words to the left of "but" is "hard, but" with the top ten variations of this theme:

1. hard, but (1,189)
2. hard but (957)
3. class, but (856)
4. class but (602)
5. tough, but (475)
6. tough but (458)
7. difficult, but (416)
8. work, but (379)
9. difficult but (334)
10. teacher, but (329)

The "difficult/synonym + but" construction, similar to what Subtirelu (2015) had found, helps commenters reconcile high difficulty and high overall ratings; e.g., "His class is hard, but he cares so much about it, and you, that you are driven to work hard for it." The "difficult/synonym + but" move predominates language around difficulty: five of its seven terms from the top 100 keywords (tough, challenging, hard, demanding, and difficult) have it as their first or second most frequent two-word cluster. This pattern suggests commenters assumed their peers may advance the leniency hypothesis by not valuing course difficulty in itself and sought to overcome such default preferences.

By closely reading full comments making use of the "difficult/synonym + but" structure, we identified general kinds of claims and the explicit or implicit warrant that support the argument (See Table 2 for examples). Commenters, for instance, readily identified various challenges, from assessment methods to subject matter. They then explained how their instructors help them to manage these difficulties. Many comments noted that, while material may be challenging, i.e., the course may be authentically difficult, HRC instructors lessen rather than aggravate these challenges. Sometimes the explanations are explicit whereas others rely on implied warrants, such as the idea that feedback, fairness, and well-prepared materials mitigate difficulty.

Table 2. Examples of HRC arguments around the "difficult/synonym + but " pattern.

Difficulty Location	**Why Difficulty Score is High**	**Why Quality Score Is High**	**Warrant**
The instructor	(a) very tough but … (b) she makes you work hard and think hard but...	(a)...very fair and very good (b) ... it is what you will have to expect in the real world	(a) Fairness and goodness outweigh difficulty (b) Real-world application makes difficulty worthwhile

The assessments	(c) Yes, his exams are hard but ... (d) He grades pretty tough but ...	(c) ... it pushes you to try harder, and if you get stuck he is more then glad to help you understand. (d)...will let you show him drafts beforehand.	(c) Assessments should reward hard work as well as (d) (d) extra support mitigates tough grading
The course subject	(e) Hard, but Chinese is hard ... (f) latin is *totally* hard ... (g) Physics is hard but ...	(e) … She is very organized and clear on her assignments and counts on verbal communication (f) … but he really likes the language. (g) … he makes it enjoyable and make SENSE?	(e-g) Recognition of *authentic difficulty*--some subjects cannot be made easy, but professors' actions and attitudes can help students overcome the challenge.

In sum, the HRC's patterns show commenters explaining their scores of high difficulty, undermining the leniency hypothesis even as they seem to assume readers will hold it. Commenters credit instructors for providing help in varied ways—course material preparation, feedback, fair assessment practices, meaningful work, and general availability—suggesting that instructors can take many paths to make their course's difficulty seem worthwhile.

Arguments Patterns in the LRC: "impossible + and"

In the LRC, the only top 100 keyword relating to difficulty is "impossible," with approximately one third of its occurrences tied to class assessments. Comments like "tests are impossible" often accompany additional criticisms: "His tests are impossible and unclear. He assigns h.w. and doesnt explain what he wants you to do." This "impossible + and" pattern recurs throughout, often with the "and" implied. LRC commenters seemed to recognize "tests are impossible" cannot adequately justify a low overall score, so they provided further evidence. We could speculate that impossible exams can be read as authentically difficult, which students might accept or at least tolerate.

LRC commenters often nested impossible exam comments within a list of artificial difficulties (Table 3). These artificial difficulties, i.e., those without direct bearing on the course's objectives, frequently included a mismatch between class expectations and exam dynamics. Comments also identify artificial difficulty when explaining class situations, such as not being able to use the bathroom or finding lectures inaudible or unrelated to tested material.

Table 3. Examples of LRC arguments around "impossible + and" pattern.

What is impossible	Why it is impossible	What else is difficult	Warrant
(a) The tests are impossible...	(a)... b/c they are far more complex than in class examples...	(a) ...Breaks promises, such as no formula sheet. She is really mean, deaf and old."	(a1) Class should prepare students for exams (a2) Class policies should be consistent and transparent (a3) Students should be treated with respect
(b) his tests are impossible ...	(b) n/a	(b) ...and he contradicts himself in his notes and tells us that what is in hte text book is wrong."	(b) Poor class presentation of information unnecessarily complicates assessments
(c)...and his exams are nearly impossible ...	(c) …to complete even with a giant curve….	(c) This guy is a flat out jerk. Inconsiderate, unhelpful,... ...His system is way to confusing to follow.	(c1) See a3 (c2) See b

One might counter that students lack expertise in test design and pedagogy. We think, however, that comments about impossible exams coupled with other issues, such as communicating expectations, often shift the difficulty from authentic to artificial. Commenters described tests in both corpora, but there is a notable difference in the phrase "tests are impossible": the LRC contains 91/m and the HRC 35/m. Furthermore, the HRC's occurrences about impossible exams are often accompany the "difficult/synonym + but" format. For example: "Class is fun, John Barr is the coolest guy, the tests are impossible but he lets you make up for them with labs and projects!" Here, the commenter acknowledged difficult exams but also provided reasons to support an overall high-quality score, namely multiple avenues for demonstrating learning. In contrast, the LRC's comments around "tests are impossible" tended to list artificial difficulties.

General Keywords in the LRC

With impossible as the only difficulty synonym in the LRC's 100 keywords, we next reviewed the top ten general keywords in the LRC. Just as the HRC's list featured the unexpected "but," the LRC includes the seemingly positive "teach." However, this word reveals why many students provided low quality and high difficulty scores; the instructor allegedly "does not teach." The most frequent two-word cluster around teach is "teach yourself," appearing 720/m times in the LRC compared to 35/m times in the HRC. Closer examination of these instances shows the "teach yourself" phrase almost exclusively refers to commenters feeling that they must teach themselves the course material; commenters appear to define teaching as extending beyond the assigning of, and lecturing on, class materials.

Longer clusters around "teach" provide further evidence that a perceived lack of teaching underpins many low LRC ratings. Among three-world clusters, "does not teach" (LRC 1,054/m, HRC

44/m) and "how to teach" (694/m, HRC 59/m) are the most common phrases, and other frequent three-word clusters reinforce this argument, such as "s/he cannot teach" (376/m LRC, HRC 6/m). Faculty and students can disagree on what teaching means, but the "does not teach" comments often highlight legitimate concerns, as for example "He does not teach the material; he just speeds through the slides while talking 100 MPH." While no speed can suit every student, issues with class pacing generally merit attention.

Other comments identify further reasonable pedagogical concerns, such as failing to model assigned tasks or explaining real-world applications. Thus, LRC commenters share similar warrants with the HRC ones, in terms of what they value; the LRC commenters just did not see their classes enacting the values. Similarly, a comment like "The professor does not teach out of the book, so there is nowhere to reference for further clarification" flags an issue with materials or communication. In fact, the most common sub-pattern within the "does not teach" line of argument focuses on a failure to communicate, such as "I can remember askign a question and his response was'just what it says,buddy.'" How instructors talk to students might be one of the elements over which they have the most control, and so the repeated comments focused on communication strike us as persuasive justifications for the LRC ratings.

Granted, in our large corpora, a handful of comments can be found to support many kinds of arguments. There are 15 comments in the LRC about faculty wearing the "same clothes," and none in the HRC. It might be tempting to draw conclusions based on that pattern, but so small a finding should be interpreted as noise. Similarly, one can find criticisms in the HRC such as "He is talented, but he cannot teach. Talks to students, as they are PhDs." These could be presented as counters to our claim that the LRC and HRC offer different commenting patterns. This would, however, obscure the issue of scale; the 1,054/m occurrences of "does not teach" in the LRC dwarfs the 44/m instances in the HRC, for instance.

We return then to the question of what allows these two corpora to exist—why are some difficult instructors highly rated overall and others not? When comparing the HRC and LRC descriptions of difficulty, we see students distinguishing between authentic and artificial types of challenges. They approve of, even applaud, the first but disapprove of and downgrade the latter. To see how these differences materialize in teacher behaviors, we next examine how the HRC and LRC varied in terms of teacher trait descriptions.

Teaching Traits Analysis

To examine how students discuss teaching traits in light of their overall quality and difficulty scoring, we drew on the Delaney et al. (2010) characteristics of effective teachers. Based on a survey of over 17,000 students, Delaney et al. find students describe effective teachers as exhibiting nine traits, listed here in descending order as to how often they appeared in the original findings: "respectful of students, knowledgeable, approachable, engaging, communicative, organized, responsive, professional, and humorous" (p. iii). These categories manifest in various ways; for example, "approachable" splits into "the positive interaction between professors and students; the comfort level of students to ask questions and to seek advice; and the sincere effort on the part of instructors to help students reach their academic goals" (p. 36).

To analyze how RMP comments evoked Delaney et al.'s (2010) findings, we created two lists of instructor traits. The first listed the nine preferred traits along with all the synonymous and linked terms defined in Delaney et al.'s (2010) report. For each trait descriptor, we searched for all appropriate instances of the word, e.g., respect, respectful, respects, etc.. Since Delaney et al.'s (2010) list of nine traits plus all of their associated descriptors proved long, we narrowed our analysis down to the three most prevalent descriptors in either the LRC or HRC. If the trait's term (e.g., respect) was not in the top three, we also included it. Thus the "respect for student" category includes the terms "help," "understand", and "care" as well as "respect" (which was not a top three word).

We found that simple word counts overlooked important commenting trends, however. This is a known limitation of corpus analysis, but we could mitigate it somewhat by reviewing the context of our narrowed list of descriptor traits. For example, we observed that "understanding" often negated "not understanding." Running a word count would include all the instances of "not understanding" towards the total of "understanding." To address this, we tallied negations of descriptions that had at least 20 occurrences and subtracted them from the trait's total occurrences.

To further identify ways students negated traits, we ran two- and three- word cluster analyses on the left and right of each of the trait's top three most prevalent descriptors. With the trait "respect," the difference between the corpora becomes clear. In the HRC, none of the top ten most frequent two-word clusters involve a negation. In the LRC, six of the top ten do. We also subtracted patterns that showed the word used to describe something other than the teaching trait. Thus, instances of "i respect" were subtracted from the trait tallies, since it concerned the commenter, not the instructor (Table 4).

Table 4. Comparison of the HRC and LRC right-side two-word clusters for respect.

HRC right-side two-word clusters for respect lemma	**LRC right-side two-word clusters for respect lemma**
i respect with respect will respect he respects of respect well respected and respect very respectful the respect much respect	no respect not respect and disrespectful very disrespectful i respect is disrespectful with respect of respect rude, disrespectful to respect

Three-word clusters showed further contextually troublesome patterns. When analyzing the use of "understand," we noted the LRC included 566 instances of "difficult to understand." After identifying the total occurrences of such two- and three-cluster negations with 20 or greater instances in either corpora, we subtracted that number from the corpus's total occurrences of the descriptor. While this will not account for all of the ways commenters reverse the meaning of descriptor traits, it provides a more accurate representation than mere frequency counts.

To contrast the use of positive instructor traits, we also developed a set of anti-traits. For example, Delaney et al. (2010) state that the trait "respect for students" links to the descriptors of "kind" and "humble." We thus created an anti-trait of "disrespectful of students" with descriptors including "mean" and "arrogant." We applied the same counting method as above, subtracting negations or irrelevant phrases (e.g., "not mean," "i mean"). Figure 3 graphs the final counts for both the traits and anti-traits.

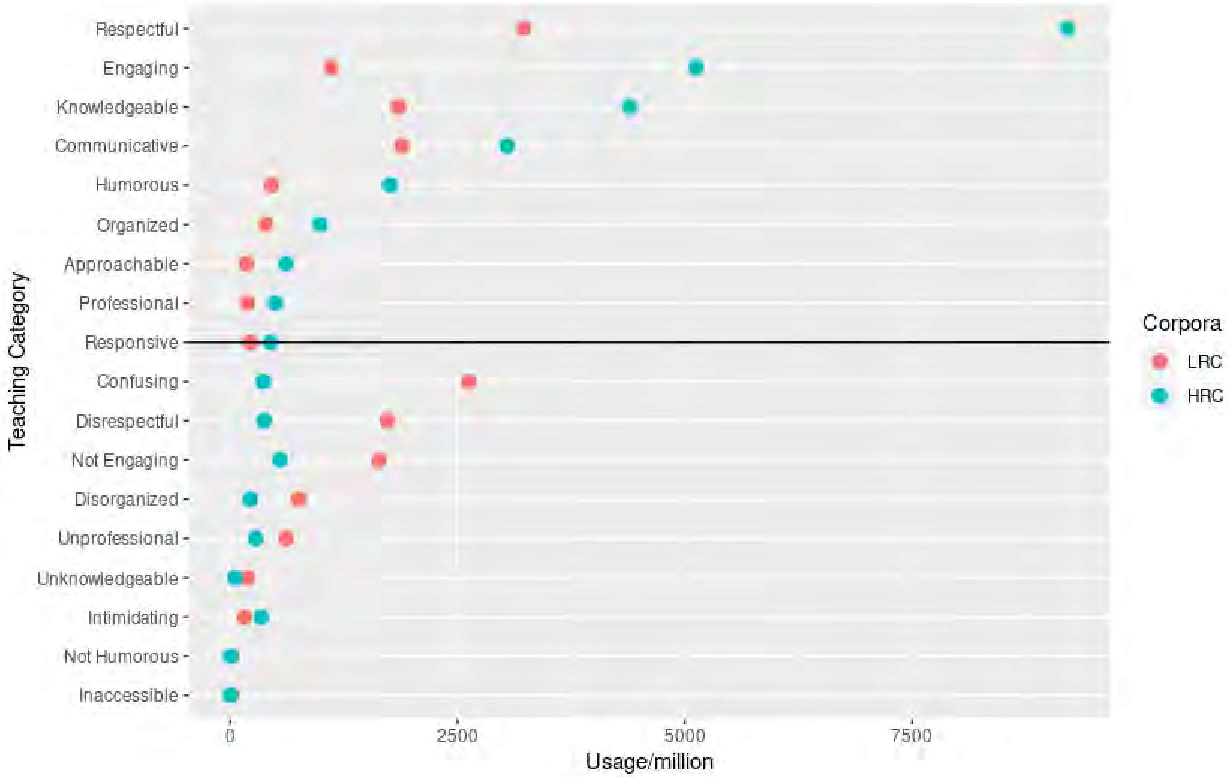

Figure 3. Teacher traits for the corpora, standardized to occurrences per million. The horizontal line separates positive from negative traits.

In each of Delaney et al.'s positive teaching traits, the HRC includes more references than the LRC, dramatically so in terms of respectfulness. Our list of anti-traits shows eight of nine negative characteristics are more prevalent in the LRC, with confusing, disrespectful, and not engaging standing out. We find it telling that respect features so prominently in the HRC, as gestures of respect are often within an instructor's control. Some gestures will be misinterpreted, of course, and some students will have inappropriate ideas about what constitutes respect. Nevertheless, the HRC's instructors appear able to convince students they are respected, a trait all faculty could aim to emulate.

We note that while we mitigated some of the limitations in corpus analysis by examining context, we could not do it for all cases. For example, when analyzing the word "responsive," we found little difference between the corpora. However, the comments revealed numerous responsive (or not) behaviors. An instructor who ignores emails or holds more office hours than required indicates something about responsiveness (or a lack thereof), but corpus analysis cannot readily capture this as neither the word "responsive" nor a synonym is used in the description. Further research could perhaps isolate a particular category such as responsiveness and closely examine commenting patterns.

Concluding Remarks

Contra Clayson and other SRI critics, RMP comments suggest students have good reasons for their judgments of difficulty and are capable of assessing instructors based on their performance rather than their mere likability. As our emotion analysis demonstrates, comments expressed starkly different overall sentiments about the corpora, despite the fact instructors in both corpora are rated highly difficult. The HRC's comments overwhelmingly represent positive emotions (trust, joy, and anticipation), whereas the LRC's comments foreground negative feelings, with fear predominating. Some instructors, RMP suggest, are able to run classes widely acknowledged as difficult and good.

Our keyword analyses offer more concrete examples of how commenters differentiated between high-quality, difficult classes and low-quality, difficult ones, specifically in comments around the concept of "difficulty." For example, the term "hard" occurred frequently in both the HRC and LRC corpora, but nuanced vocabulary (challenging, demanding, difficult) appeared more often in the HRC. One reason for this split was because commenters in the HRC attempted to explain why a difficult class/instructor remained effective. Commenters sometimes struggled to articulate what made such classes worthwhile—perhaps because they were novice evaluators and writers, perhaps because they did not put forth much effort—but they clearly felt that the HRC classes were valuable. Difficulty proved insufficient in itself to deter high quality ratings.

The prevalent, contrasting patterns of HRC "difficult/synonym + but" and LRC "impossible + and" illustrate how RMP commenters justified ratings in terms of between authentic and artificial difficulty. Commenters described how HRC faculty provide strategies for managing difficult work, including "impossible tests." HRC comments praised relevant lectures, helpful materials, feedback and review sessions, and general availability. These features overcome a perception of artificial difficulty without eliminating difficult exams. Meanwhile, the LRC corpus suggests that implementing artificial difficulties can lead students to view not just exams as impossible but rather the overall class as unnecessarily hard. Yet as the "impossible + and" pattern suggests, it is a combination of difficult assessments along with more artificially imposed hardships that earn LRC-level quality scores. Commenters repeatedly criticized LRC instructors for being "unhelpful," "confusing," and for "not communicating" expectations or class material; traits that seem to name concrete teaching actions rather than just focus on personality. Similarly, in the LRC, the phrase "did/does not teach" materialized often, alongside synonymous ideas like having to "teach yourself."

These contrasting patterns do suggest that commenters view difficulty, unqualified, as a negative feature of the class. In this way, we see them sharing a general assumption with SRI deniers and with Chaney (2011) specifically: students may not value class "difficulty" in and of itself, and commenters very well might, as Chaney (2011) argues, approach RMP "to have their consumer needs met" (p. 197). But, whereas Chaney (2011) claims "most positive instructor characteristics (from humor, to knowledgeableness, to verbal skill) are the means to the most desirable end of ease" (p. 198), our analysis finds HRC commenters arguing that these kinds of traits exist alongside elements of difficulty; their presence serves to make classes worthwhile rather than merely easy.

Chaney (2010) also argues that complaints of disorganization, irresponsibility, and even the charge of losing student papers, represent:

> Extreme impatience for any evidence of the professor's humanity that might still be infecting the classroom. The very fact that this complaint is so typical in the students' rhetoric suggests that it is also reflective of the values around which this public organizes. (p. 199)

We agree that RMP commenters seem to share values, but we see them as aligned with traits of good teaching rather than consumerism. A day or two of disorganization is human, but consistently arriving unprepared, failing to connect lectures to class assessments, and losing student work seem to us less momentary lapses and more systematic evidence of teaching malpractice. And it is the latter we see commenters emphasizing in the LRC. Having an instructor lose your paper might be a lesson in how powerful systems fail to treat individuals' paperwork with care. But learning to negotiate paperwork is not usually a stated objective of college courses. It looks instead like a rather artificial difficulty—a professor saying "you finished your work as assigned, but you also need a backup plan in case I don't keep it safe."

Our analysis suggests that commenters themselves do not solely prioritize ease. Rather, we see commenters making a more nuanced evaluation of difficulty than is often assumed by SRI scholarship. RMP participants differentiate between authentic and artificial forms of hardship, appreciating the former and denouncing the latter. Since the current form of RMP provides only one way to rate difficulty, students turn to the comments to explain their evaluation and define what kind of difficulty other students can expect. Thus, when SRI deniers argue that SRIs punish difficult instructors, they do not appear to understand the different ways in which difficulty can be defined.

This recalls Marsh's (2001) concept of "good" and "bad" workloads; there appears to be "good" and "bad" difficulty. Rather than call difficulty good or bad, though, we prefer the terms authentic and artificial. The leniency hypothesis implies that certain course material is inherently "difficult" and that those who teach it are punished unfairly. This seems to us a mistake. Comment patterns suggest that difficulty ratings arise not only from the course material itself but also from the instructor's handling of it. RMP commenters, our findings show, often enjoy challenges so long as the material is being taught and not merely presented.

On the Perennial Question of Gender and Racial Bias

No discussion of SRI literature can avoid questions about other suspected biasing factors, especially those of gender and race/ethnicity. Research on gender bias in particular features a long history and near constant updates, but the findings are ambiguous. A full review of this particular debate is beyond the scope of our article; for interested readers about the effects of gender bias on SRI results, we recommend Li and Benton's (2017) analysis of over 25,000 instructors using a validated evaluation instrument, and for those interested in an overview of SRI in general we recommend the Benton and Ryalls (2016) and Wang and Williamson (2022) reviews.

As described in our corpus demographics section, we did not see a marked difference in the corpora demographics we analyzed. Since our analysis aimed only at checking for substantial differences, we caution against reading our demographic results as evidence for or against particular student biases. For instance, the same percentage of Black-identified names appeared in both corpora (7.7%). Would we argue this refutes claims of student racial bias affecting SRI? No. We do not think our study allows for such inferences as we built no models and only reported on this one statistic. Similarly, we can conclude that female-identified faculty were slightly more likely to be found in the LRC compared to both the HRC and a sampling of RMP instructors without inclusion criteria. The difference, however, was quite small. We thus reject arguments that our corpora offer sufficient evidence to claim student gender bias affects RMP rating, as such work would need to be modeled.

Even for readers who see the difference in gender representation as meaningful, we cannot assume the difference was due to students' bias. Other course elements, such as required versus elective status and upper versus lower level have consistently been shown to affect SRIs (see research summaries by Marsh, 2007; Benton & Ryalls, 2016). It could be the female-identified faculty in our sample taught more lower-level, required courses and their scores reflect students' biases against mandated, intro classes. While this kind of course assignment might reflect gender bias, it is not one located in students. Further research could examine how course type interacts with students' perceptions of difficulty and quality.

A marked difference in our corpora is that Asian/Pacific Islander (API) names were more likely to be in the LRC. It is tempting to draw conclusions that these results show racial bias in SRI scores against API faculty. This finding would be in keeping with similar results elsewhere (Subtirelu, 2015). Again, though, our data cannot speak as to why. These racial data are, for example, confounded by discipline. STEM fields were overrepresented in the LRC; the top 5 disciplines were mathematics, English, computer science, engineering, and what RMP labeled "Science." Outside of English, these are disciplines in which API faculty may be overrepresented; the National Center for Education Statistics (2020) estimated, for example, that in 2018, API comprised 12% of all postsecondary faculty, but the American Society for Engineering Education estimated Engineering faculty to be 28% API (2020). Moreover, Census data do not consider the different cultures and ethnicities captured by the term API. These data are far too preliminary to draw any clear conclusions. Further research could probe the potential for racial bias in RMP comments as they pertain to difficulty.

General Implications for Researchers and Teachers

Our overall findings have both theoretical and practical implications. Many proponents of the leniency hypothesis argue that students cannot rate instructors because students do not understand good teaching. Our findings suggest, instead, that students might have a much better sense of what constitutes effective teaching than research-based faculty. College students are obviously exposed to various teaching contexts: instructors who engage, those who present, and all manner in between. Students, moreover, have significant experience in contemporary classrooms, as many have sat in them for over a decade by the time they arrive at their higher educational institutions. And they are the primary audience of any class lesson. A priori, we should expect students to have a firm grasp of what constitutes effective teaching and should take heed of their views.

While our findings lend support to the validity of SRIs, we are not arguing that SRIs should be the sole or even primary form of faculty evaluation. Best practices for faculty evaluation emphasize that SRIs represent only one form of data, to be used alongside other measures of teaching quality. Building off those core principles, this analysis suggests faculty evaluators (and faculty engaged in self-reflection) should be cautious in how they interpret quantitative scores of “difficulty.” Without qualitative feedback to clarify the type of difficulty being rated, such scores are ambiguous.

In parallel, these findings suggest ways of reading comment patterns. Individual comments, we believe, rarely merit attention from faculty evaluators. Anyone who has taught even a handful of semesters can point to an inappropriate or unjustified comment; we might acknowledge praise is sometimes also hyperbolic. Patterns of response, however, do provide useful information. Repeated claims of “not teaching” likely flag a serious pedagogical issue; feedback along the “difficult + but” pattern suggest the instructor is developing authentically difficult courses. We stress the importance of patterns—repeated occurrences of such responses help define a faculty member’s weakness or strength; a single comment does not.

We would like to end by anticipating the rebuttal that more instructors are rated highly in both quality and ease than are rated high in quality and in difficulty, possibly supporting the leniency hypothesis. First, we observe that teaching effectiveness rests not on difficulty per se. Instructors might indeed create a sense of ease through effective teaching, especially if their good practices make learning less work than expected. We also note our findings show raters view difficulty as something beyond the mere absence of ease, and so arguments about what students mean by “easy” do not necessarily reveal what they mean by “difficult”.

That said, we agree guiding students through authentically difficult work is challenging; instructors that manage to do so may be relatively scarce. It does not hold that, because students only see some faculty as providing excellent, difficult classes, that SRI are invalid. Indeed, we think such findings could be just as easily interpreted as evidence of students’ discernment and thus would support the overall validity of SRI. While Clayson and Sheffet argue that “attempts to produce master teachers would be a waste of precious time” (p. 159), due to what they believe to be a confounding halo effect, we disagree. Teaching is a difficult profession; masters of the art are rare. But that is not reason enough to forego the challenge. It is in attempting authentically difficult work, after all, that we truly learn.

Appendix

Appendix 1. Method Notes

We chose 2,300 instructors for each group as 2,300 was, for us, a computationally feasible value. Our script kept crashing (timing out) with high instructor numbers, and on our computers the corpus analysis software struggled to open large text files. We found 2,000 a good number and then incremented by 1,000 until finding a value which did not crash our software. The final number of 4,600 seemed a fair balance between computation/analysis memory limits and comprehensiveness.

To collect data, we wrote a Python program (see the below scraper) to parse RMP scores and capture the comments for the instructors that met the inclusion criteria.

Corpus analysis method notes

Both corpora are freely available [in files submitted with manuscript]. All RMP comments included here are reproduced in their original form, including typos.

In corpus analysis, the "keyness" of a word is a test statistic, which here is the log likelihood. The higher the "keyness" of a word, then the less likely it is to have occurred by chance. If "keyness" bypassed the software's most stringent significance threshold (p < 0.0001), then we considered it significant.

We chose keywords which were significant but also sensible, following a recommendation to consider significance testing as an exploratory means to identify words worthy of further study, rather than as a confirmatory tool (Bestgen, 2014). Keyness suffers from the same limitations of any significance test; for example, slight differences become significant in large sample sizes. We, therefore, only reported significant keywords which were clearly interpretable and also had large raw frequency differences. For example, we reported "boring" not solely because of its keyness but because it made contextual sense and had a large frequency difference (373/m in the HRC compared to 1,337/m in the LRC).

If you would like to view the complete code sequence and coding notes, please contact the authors.

References

Anthony, L. (2020). *AntConc* (3.5.9) [Computer software]. http://www.laurenceanthony.net/software/antconc/

ASSE, A. S. for E. E. (2020). *Engineering & Engineering Technology By the Numbers.* https://ira.asee.org/by-the-numbers/engineering-faculty/

Benton, S. L., Guo, M., Li, D., & Gross, A. (2013). Student ratings, teacher standards, and critical thinking skills. *Annual Meeting of the American Educational Research Association, San Francisco.*

Benton, S. L., & Ryalls, K. (2016). Challenging misconceptions about student ratings of instruction. *The IDEA Center*, *IDEA Paper 58*.

Bestgen, Y. (2014). Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing 29*(2), 164-170. https://doi.org/10.1093/llc/fqt020

Bleske-Rechek, A., & Michels, K. (2010). RateMyProfessors com: Testing assumptions about student use and misuse. *Practical Assessment, Research, and Evaluation*, *15*(1), 5. https://doi.org/10.7275/ax6d-qa78

Brockx, B., Van Roy, K., & Mortelmans, D. (2012). The student as a commentator: Students' comments in student evaluations of teaching. *Procedia - Social and Behavioral Sciences*, *69*, 1122–1133. https://doi.org/10.1016/j.sbspro.2012.12.042

Carlozzi, M. (2018). Rate my attitude: Research agendas and RateMyProfessor scores. *Assessment & Evaluation in Higher Education*, *43*(3), 359–368. https://doi.org/doi.org/10.1080/02602938.2017.1348465

Chaney, S. B. (2011). Rankings and ravings in the academic public. *Rhetoric Review*, *30*(2), 191–207. https://doi.org/10.1080/07350198.2011.552381

Clayson, D. E. (2014). What does Ratemyprofessors.com actually rate? *Assessment & Evaluation in Higher Education*, *39*(6), 678–698. https://doi.org/10.1080/02602938.2013.861384

Clayson, D. E., Frost, T. F., & Sheffet, M. J. (2006). Grades and the student evaluation of instruction: A test of the reciprocity effect. *Academy of Management Learning & Education*, *5*(1), 52–65. https://doi.org/10.5465/amle.2006.20388384

Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education*, *28*(2), 149–160. https://doi.org/10.1177/0273475306288402

Coladarci, T., & Kornfield, I. (2019). RateMyProfessors.com versus formal in-class student evaluations of teaching. *Practical Assessment, Research, and Evaluation*, *12*(1). https://doi.org/10.7275/26ke-yz55

Constand, R. L., Pace, R. D., & Clarke, N. (2016). Accounting faculty teaching ratings: Are they lower because accounting classes are more difficult? *Journal of Accounting and Finance; West Palm Beach*, *16*(4), 70–86.

Dayton, A. E. (2015). Making sense (and making use) of student evaluations. In *Assessing the Teaching of Writing: Twenty-First Century Trends and Technologies* (pp. 31–44). UP Colorado.

Delaney, J., Johnson, A., Johnson, T., & Treslan, D. (2010). *Students' perceptions of effective teaching in higher education* (p. 90). St. John's, NL: Distance Education and Learning Technologies. https://research.library.mun.ca/8370/1/SPETHE_Final_Report.pdf

Hoyt, D. P., & Lee, E.-J. (2002). *Basic data for the revised IDEA system* (Technical Report No. 12; p. 87). Individual Development and Educational Assessment. https://www.ferris.edu/administration/academicaffairs/Resources/CourseEval/Administration/Documents/IDEA12_techreport_2002.pdf

Kušen, E., Cascavilla, G., Figl, K., Conti, M., & Strembeck, M. (2017). Identifying emotions in social media: Comparison of word-emotion lexicons. *2017 5th International Conference on Future*

Internet of Things and Cloud Workshops (FiCloudW), 132–137. https://doi.org/10.1109/FiCloudW.2017.75

Lauer, C. (2012). A comparison of faculty and student perspectives on course evaluation terminology. *To Improve the Academy*, *31*(1), 194–211. https://doi.org/10.1002/j.2334-4822.2012.tb00682.x

Li, D., & Benton, S. L. (2017). The effects of instructor gender and discipline group on student ratings of instruction: IDEA Research Report #10. In *IDEA Center, Inc.* IDEA Center, Inc. https://eric.ed.gov/?id=ED588355

Marks, N. B., & O'Connell, R. T. (2003). Using statistical control charts to analyze data from student evaluations of teaching. *Decision Sciences Journal of Innovative Education*, *1*(2), 259–272. https://doi.org/10.1111/j.1540-4609.2003.00020.

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on students' evaluations of teaching. *American Educational Research Journal*, *38*(1), 183–212. https://doi.org/10.3102/00028312038001183.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (pp. 319–383). Springer Netherlands. https://doi.org/10.1007/1-4020-5742-3_9.

Otto, J., Sanford Jr, D. A., & Ross, D. N. (2008). Does Ratemyprofessor.com really rate my professor? *Assessment & Evaluation in Higher Education*, *33*(4), 355–368. https://doi.org/10.1080/02602930701293405.

Ritter, K. (2008). E-valuating learning: Ratemyprofessors and public rhetorics of pedagogy. *Rhetoric Review*, *27*(3), 259–280. https://doi.org/10.1080/07350190802126177.

Santamaria, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science, 4*. https://doi.org/10.7717/peerj-cs.156.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching the state of the art. *Review of Educational Research*, *83*(4), 598–642. https://doi.org/10.3102/0034654313496870.

Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis. *Basic and Applied Social Psychology*, *42*(4), 276–294. https://doi.org/10.1080/01973533.2020.1756817.

Subtirelu, N. C. (2015). "She does have an accent but…": Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com. *Language in Society*, *44*(1), 35–62. https://doi.org/10.1017/S0047404514000736.

Teagarden, A., & Carlozzi, M. (2020). Analyzing Student Evaluations of Teaching: A Generic Prescription. *WPA: Writing Program Administration-Journal of the Council of Writing Program Administrators*, *43*(2).

Timmerman, T. (2008). On the validity of RateMyProfessors.com. *Journal of Education for Business*, *84*(1), 55–61.

Wang, G., & Williamson, A. (2022). Course evaluation scores: Valid measures for teaching effectiveness or rewards for lenient grading? *Teaching in Higher Education*, *27*(3), 297–318. https://doi.org/10.1080/13562517.2020.1722992.