Lexical Measurement of Teaching Qualities

Laura Biester
Ian Stewart
Dr. Laura Hirshfield, University of Michigan

Laura Hirshfield is a Diversity, Equity, and Inclusion lecturer and research assistant at the University of Michigan. She received her B.S. from the University of Michigan and her Ph.D. from Purdue University, both in chemical engineering.

Rada Mihalcea
Sara Pozzi

1 Introduction

It is common practice to utilize course evaluations to have students anonymously rate their instructor's teaching ability, and other aspects of the course experience. These evaluations tend to include both numerical (Likert scale) and open-ended written feedback, although thorough analyses of written feedback are rare due to the lack of methods to rigorously analyze the large amount of content with a teaching-specific lens. In this paper, we create a comprehensive lexicon to measure eight teaching qualities from the written feedback using a combination of natural language processing (NLP) and manual filtering. We refer to this lexicon as "Lexicon for Evaluation of Education Quality" (LEEQ). We then validate LEEQ by analyzing how the frequency of words in each dimension is correlated with (a) numerical ratings and (b) other dimensions. Finally, we compare it with other sentiment analysis tools that are less fine-grained, as overall sentiment scores may not capture teaching-related qualities and do not differentiate between fine-grained teaching qualities such as helpfulness and clarity.

LEEQ can be used by the research community to allow for full analyses of teaching evaluations, rather than focusing solely on quantitative metrics; in this paper, we perform a case study that highlights one such analysis. Prior work has found that course evaluations can easily be biased against certain identity groups; for example, female instructors and instructors of color tend to be rated lower or more harshly compared to white male instructors [1, 2]. The switch from traditional in-person learning to hybrid or remote learning during the COVID-19 pandemic also likely influenced student perceptions of their educational experience. Our case study uses the lexicon as a lens to answer the following research questions: what differences occur in free-text course evaluations between in-person and remote/hybrid learning? Further, what differences arise between instructors of different identities?

To address these research questions, we analyzed all course evaluations submitted for College of Engineering courses at a large Midwestern institution, from six semesters spanning Winter 2019 to Fall 2021. Rather than considering the numerical ratings, we aim to more closely examine the student comments, to determine if any biases arise in how students describe their instructors or course experience. We find changes in the frequency of words representing high and low quality instruction, and find that students refer to instructor's helpfulness more often during COVID-19. We do not find significant differences in the frequency of words representing teaching qualities based on the sex of the instructor, either before or during the pandemic. In all, the results signal that a shift to a remote format does not have negative implications on the language used to describe instructors in their evaluations.

2 Related Work

2.1 NLP in Education

As educators increasingly use technology to improve course delivery, researchers have begun to test a variety of machine learning methods to help educators and students navigate their courses more effectively [3]. On the student side, chat bots powered by NLP can help students by intelligently retrieving requested information and suggesting related topics for students to explore [4, 5]. From the educator's side, NLP systems can process student writing at scale e.g., to detect plagiarism [6] and give feedback on grammatical errors [7]. In the domain of written evaluations, researchers have developed NLP models to detect unfair descriptions of female professionals, which includes identifying linguistic stereotypes within documents and within

word choices [8, 9]. While more complicated to analyze than numerical responses to a survey, written text provides useful signals from students that would otherwise be unavailable, such as a instructor's engagement with students during lectures [10].

2.2 Bias in Instructor Evaluations

Prior work has established a consistent bias in the evaluations that college students provide for their instructors, particularly based on gender, race, and ethnicity (e.g., harsher reviews for underrepresented instructors) [1, 11, 12]. Bias in evaluations is prevalent on public websites such as Rate My Professors, where students' responses are anonymized and can potentially reach a wide audience of fellow students [13, 14]. The degree of bias expressed may depend on how well a minority group is represented in higher education [15] or whether students have differing expectations [16]. The bias may take the form of explicit differences in ratings, e.g., rating male instructors as better teachers [2], or implicit differences in language use, e.g., using abusive language to describe an instructor [17]. In addition to the personal harm done to instructors, bias can derail the careers of minority-group instructors as course evaluations often play an large role in determining tenure and promotion [18, 19]. Our case study builds on the well-established notion of bias in student evaluations, and we investigate how much bias exists in written evaluations and whether that bias changed when courses switched to virtual format in 2020.

3 Methods

3.1 Data Collection

Our new data set, henceforth CCE for "COVID-19 Course Evaluations," comes from a public university in the U.S. Midwest. The university's registrar provided 23,882 course evaluations from the College of Engineering collected over six semesters, from Winter 2019 to Fall 2021 (Table 1). We analyze data based on instructor sex and COVID; future work will analyze this data for other demographic characteristics, such as race/ethnicity and nationality.

Data pre-processing Some of the courses in our data are associated with multiple instructors, and some of the evaluations target teaching assistants rather than professors. We perform extensive data cleaning before analysis, including: (1) removing multi-instructor courses (courses where multiple instructors were referenced in the evaluations); (2) removing lab and recitation sections where teaching assistants were likely to be mentioned in evaluations; (3) replacing personal names detected with Named Entity Recognition[1] with generic `PERSON` tokens; and (4) removing short (less than 10 tokens) and duplicate evaluations.

This filtering process reduced our data by 33%, yielding a total of 16,010 course evaluations for analysis. 1,075 instructors (summed across pre and post-COVID periods), 651 courses, and 35 departments (including cross-listed departments/courses outside of engineering) are included in our filtered data, and evaluations have on average 45.4 tokens.

For most of the evaluations, students were required to write a response to at least one question related to teaching quality. We merge three separate questions that were phrased differently for different courses but address the same core concerns of perceived teaching ability.[2]

[1] Using the default NER system in Spacy: `https://spacy.io/api/architectures/#parser`

[2] "Comment on the quality of instruction in this course," "Please comment on the effectiveness of the instructor," and "Please enter any additional comments you have for the instructor."

Demographic	Total	Percent
Sex		
Male	812	76.0
Female	263	24.0
Ethnicity		
White (Not of Hispanic Origin)	656	61.7
Asian	243	22.1
Not Indicated	96	8.5
Black/African American	29	2.7
Two or More Races	27	2.6
Hispanic/Latino	24	2.4
Citizenship		
United States	780	72.6
China	81	7.6
India	45	4.3
Other	169	15.6

Table 1: Instructor demographics (after filtering). Identity groups with fewer than n = 10 unique instructors per-group per-period are grouped in the "Other" category for statistical and anonymity purposes. The values are summed across pre and post-COVID periods.

3.2 Lexicon Construction

We develop a lexicon to understand the text of course evaluations, which we refer to as Lexicon for Evaluation of Education Quality (LEEQ). The lexicon capture eight paired teaching qualities: **high quality/low quality**, **helpful/unhelpful**, **easy/difficult**, and **clear/unclear**. We selected these qualities due to the range of features of instruction that they cover, in addition to the availability of paired numerical ratings and text associated with each of them (as described in the following paragraph). We provide examples of words from the lexicon's eight dimensions and example sentences in Table 2. Throughout the paper, we will denote the names of the lexicon dimensions in **bold**, while individual words from the dimension are underlined in examples.

These dimensions cover a wide range of desirable and undesirable teaching behaviors, including lecture comprehensibility, student feedback capabilities, adaptability to different needs, and social support provided to students. Furthermore, we built the lexicon so that it would not be inherently linked to different instructor's social groups, e.g., the lexicon does not explicitly refer to gender. We developed the lexicon through a large-scale analysis of a separate data set from the popular Rate My Professors (RMP) website, containing 863,857 total reviews from 2015 and earlier across 31 universities in North America, drawn from [20].[3] RMP hosts public reviews for professors written by students from across U.S. institutions, and it includes both text comments and scores for various categories, including quality, helpfulness, difficulty, and clarity.[4]

We aim to identify individual words that are indicative of the different aspects of teaching quality

[3] `https://www.ratemyprofessors.com/`. Canadian universities with a large number of French reviews are excluded from our analysis.

[4] We exclude some ratings that are available on the website such as the student's prior interest in the course, as these are less connected to instruction. A screenshot of the ratings form from 2014 is shown in Appendix A.

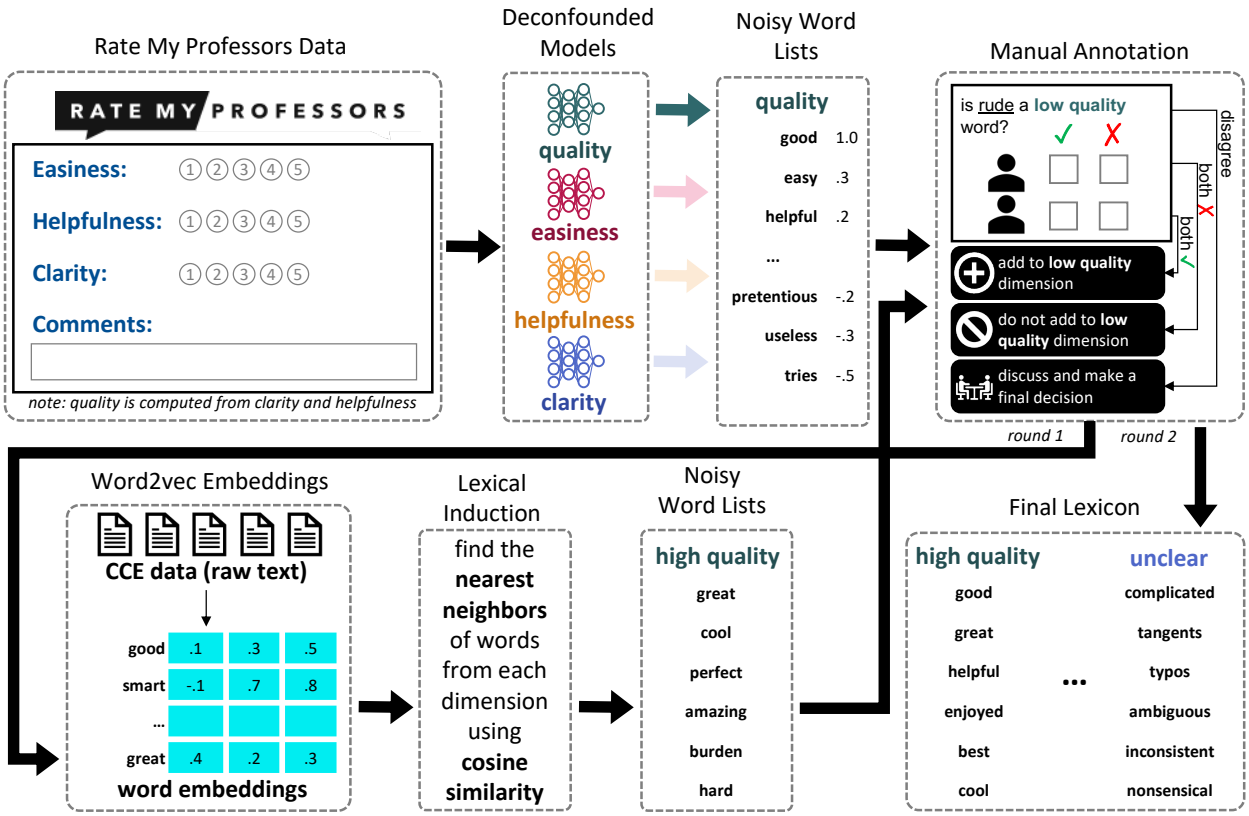

Figure 1: Diagram representing the development of our lexicon.

specified above. Identifying words in the RMP data that are correlated with these teaching aspects, e.g., words associated with high "difficulty" ratings, can result in substantial overlap between dimensions due to correlations between the categories. Many instructors who receive high "difficulty" ratings are likely to receive low "quality" ratings as well, and therefore the words correlated with "difficulty" are also likely to be correlated with low "quality." We handle this problem with a method for de-confounding correlated lexicons [21] that trains a deep learning model to score words more highly based on their cooccurrence with a single category and non-cooccurrence with the other confounding categories. We use sub-sets of the other dimensions and metadata, e.g., course department, as confounders for model to remove. Using this method, for the **difficult** dimension, the model learns to identify words that are more correlated with higher difficulty ratings but not correlated with quality ratings.

From these word lists, two of the authors manually annotated the words that were valid members of the different dimensions based on fixed criteria. For example, the word "helping" would count as **helpful** but not **clear** because **helpful** words should reflect positive social behavior while **clear** words indicate effective communication. Next, we adapt these lists to the original CCE dataset by computing the nearest neighbors to the words in each dimension, using word embeddings trained on the CCE text data [22]. Computing the nearest neighbors to a word such as "helping" reveals similar words such as "aiding" and "guiding" which the original lexicon did not identify from the Rate My Professors data. After filtering the expanded lists through another round of manual

Dimension	Unique words	Freq. (per 1K words)	Most frequent words	Example sentence	κ_{OD}	κ_{LI}
High quality: positive student experience	165	28.9	good, great, helpful, enjoyed, best	Professor PERSON was an amazing instructor. Hands down the best professor I've had so far at the university.	0.63	0.88
Low quality: negative student experience	212	4.15	confusing, unclear, confused, frustrating, poor	Synchronous class time was riddled with mistakes which lead to a frustrating experience.	0.45	0.78
Easy: coursework easy to complete	42	1.8	easy, understandable, simple, easily, basic	Professor PERSON knows what he's talking about and is able to make information easily understood	0.72	0.55
Difficult: coursework difficult to complete	72	5.5	difficult, hard, fast, challenging, difficulty	His style of challenging exams that are typically a time crunch I thought was not quite as effective.	0.64	0.35
Helpful: pro-social behavior	65	12.2	helpful, help, helped, nice, feedback	He was always pushing the minds of the students to understand the real world of engineering, and was there to help through any struggles or questions regarding course materials and assignments.	0.52	0.59
Unhelpful: anti-social behavior	47	0.5	unfair, rude, condescending, disrespectful, unreasonable	Prof. PERSON was often condescending and rude.	0.25	0.56
Clear: clear explanation of course content	40	6.9	clear, clearly, engaging, explanations, organized	PERSON is a great professor, and always gave clear and concise answers to questions.	N/A	0.59
Unclear: unclear explanation of course content	19	0.4	complicated, tangents, typos, ambiguous, inconsistent	Super inconsistent exam and HW quality/expectations that made it hard to know how to prepare.	0.72	0.75

Table 2: Summary of lexicon to analyze teaching evaluations. The right side of the table presents Cohen's κ for manual annotations of inclusion in lexicon (κ_{OD} refers to the original deconfounded lexicon, while κ_{LI} refers to items added by lexical induction). Cohen's κ cannot be computed for the original **clear** dimension because it was jointly annotated by both annotators.

annotation, we arrive at a teaching-related lexicon with a range of 19-212 words per dimension.[5]

The **clear** dimension was jointly annotated for the original words by both annotators to finalize the annotation process. The annotators discussed each word for which there was a disagreement to determine the final label. We measured inter-annotator agreement using Cohen's κ [23]; the results are shown in Table 2. Agreement ranged from 0.25–0.88. The low agreement score for the **unhelpful** dimension relates to an initial disagreement between the annotators regarding what that dimension represented. After agreeing on the definition "socially positive, willing to connect to students," the agreement score for the **unhelpful** dimension increased to 0.56 when the same two annotators labeled word pairs from the lexical induction method. Agreement increased across most dimensions during the second round of labeling. The full process for annotating our lexicon is shown in Figure 1.

3.3 Lexicon Correlations

We confirm that the lexicon is a valid construct for student assessment of teachers in the CCE dataset by examining each dimension's Spearman correlation with student's numerical ratings of

[5]The lexicon is available online: `https://github.com/MichiganNLP/LEEQLexicon`.

Dimension	Correlation
High quality	**0.3941**
Helpful	**0.0822**
Clear	0.0086
Easy	**-0.0685**
Difficult	**-0.2754**
Low quality	**-0.4281**
Unhelpful	**-0.2292**
Unclear	**-0.2184**

Table 3: Correlation between lexicon frequency and ratings of teaching quality, computed per-instructor using mean scores. Dimensions are ordered such that positive teaching qualities are first, followed by difficulty level and negative teaching qualities. **Bold** indicates $p < 0.05$ with FDR adjustment for multiple comparisons.

Method	Correlation
Vader	**0.5234**
LIWC	**0.4119**
LIWC (zscore)	**0.4209**
TextBlob	**0.4545**
Flair	**0.6387**
LEEQ	**0.4297**
LEEQ (zscore)	**0.4736**
LEEQ (Vader)	**0.5047**
LEEQ (Vader boosted)	**0.5561**

Table 4: Correlation between scores given by existing sentiment analysis tools and teaching quality ratings, computed per-instructor/semester using mean scores. **Bold** indicates $p < 0.05$ with FDR adjustment for multiple comparisons.

teaching quality and with one another (average ratings across the dataset were 1.29 ± 1.02 on a Likert scale from -2 to 2). We also compare these correlations with correlations between existing sentiment analysis methods and instructor ratings. We indicate results that are significant ($p < 0.05$) following an adjustment for false discovery rate (FDR) [24] with $\alpha = 0.05$, which adjusts the p-values to account for multiple comparisons. The adjustment is performed at the per-table/figure level.

We compute the correlation between the mean lexicon frequency per-instructor and the mean numerical rating per-instructor for overall teaching quality. As expected, the dimensions with a positive orientation (e.g., **high quality**) have a strong positive correlation with high teacher ratings, while the dimensions with a negative orientation (e.g., **low quality**) have a negative correlation. We show all correlations in Table 3.3. While the **easy** and **clear** dimensions do not have strong correlations with the teaching quality rating score, this may relate to the fact that these dimensions relate both to teaching performance and to the course itself (e.g., "homework was easy"). Students may also mention **easy** words as a preface to more negative comments in their evaluations (e.g., "the course was easy but PERSON was unpleasant").

As a further test of validity, we compute correlations between pairs of dimensions, using mean of lexicon frequencies per-instructor (Figure 2). We expect that teaching qualities that are generally positive (**high quality**, **helpful**, **clear**) should be positively correlated with one another, while teaching qualities that are generally negative (**low quality**, **unhelpful**, **unclear**) should be negatively correlated with one another. We find that this is the case; however, there are positive correlations between some positive and negative teaching qualities. We expect that this may be linked to the presence of negations in evaluations (e.g., "not difficult"), a limitation that is discussed in Section 5.

Finally, we ask to what extent existing NLP methods capture student's ratings of instructors. We

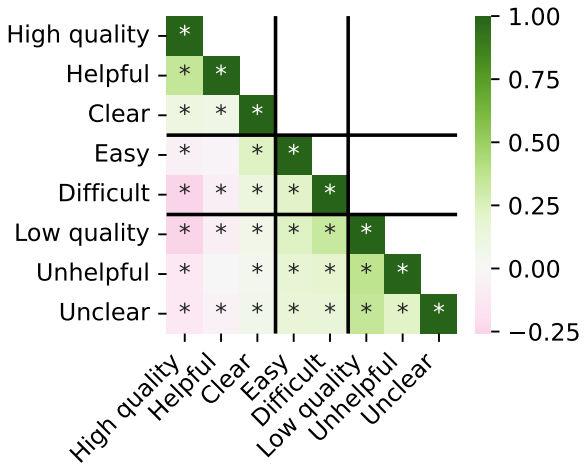

Figure 2: Heatmap representing the correlation between the lexical frequency of pairs of dimensions (e.g., frequency of difficult and unhelpful words), computed per-instructor/semester using mean scores. Color represents the correlation coefficient, while * indicates $p < 0.05$ after FDR adjustment. Dimensions are ordered such that positive teaching qualities are first, followed by difficulty level and negative teaching qualities.

compute the correlations between sentiment scores from existing methods and instructor's ratings, shown in Table 3.3. We use the following existing methods:

Vader is a rule-based sentiment analysis model [25] that uses a weighted sentiment lexicon (with positive and negative polarity for each word) in combination with context-based heuristics, e.g., the use of negation that will reverse the sentiment of a given word, or the punctuation such as !!! that heightens the existing sentiment.

LIWC The Linguistic Inquiry and Word Count lexicon (LIWC) [26] provides two scores for positive emotion and negative emotion. As a baseline, we subtract the raw negative emotion score from the positive emotion score. Additionally, we normalize positive and negative emotion across all evaluations, take the z-score, then take the difference, resulting in the LIWC (zscore) correlation.

TextBlob is a Python NLP library that provides a number of text analysis tools, including sentiment analysis.[6]

Flair provides a sentiment model that was trained on a movie review dataset using deep learning [27].

LEEQ is *complementary* to these existing methods. While it will not capture emotion in text to the same extent as other methods, it does capture teaching-specific qualities. Therefore, we do not expect it to yield higher correlations than sentiment methods on its own. It can be used in tandem with existing methods that overlook teaching-specific words to predict ratings, but perhaps more importantly, the lexicon *goes beyond overall teaching quality or polarity* in the text to identify mentions of specific teaching qualities like clarity and helpfulness.

[6] `https://textblob.readthedocs.io/en/dev/index.html`

We use the same baseline methods for our LEEQ lexicon as we use for LIWC, utilizing the **high quality** and **low quality** dimensions. We also experiment with using our lexicon with the Vader sentiment analysis method's heuristics. Finally, the LEEQ (Vader boosted) method "boosts" the weights of terms in our **high quality** and **low quality** dimensions in the Vader lexicon, yielding the highest correlation among lexicon-based methods.

We find that our lexicon's correlations with teaching ratings do not exceed all existing methods. However, they outperform purely lexicon-based approaches such as LIWC, and perform well in ensemble settings (e.g., LEEQ (Vader boosted)). The fact that our lexicon is specific to teaching qualities allows it to yield a higher correlation with instructor ratings than LIWC, even though the LIWC lexicon has higher coverage on the CCE dataset (55.7 per 1000 words in the PosEmo dimension and 12.5 per 1000 words in the NegEmo dimension). When considering the example sentences in Table 2, we find that the existing sentiment methods tend to correctly predict polarity. The main exception is the sentence "super inconsistent exam and HW quality/expectations that made it hard to know how to prepare," for which only Flair correctly identified negative sentiment.

4 Case Study

As a case study, we use LEEQ to compare the language used in evaluations in the CCE dataset prior to and during the COVID-19 pandemic. In addition to considering general differences in evaluations during these two periods, we ask whether there is a sex difference in evaluations both before and during COVID, and if so, in which teaching-related linguistic dimensions these differences occur. Our findings can shed light on differences in evaluations during the pandemic, as well as between in-person and remote teaching.

As discussed in Section 3.1, the CCE dataset includes evaluations from Winter 2019 to Fall 2021. For our case study, we drop evaluations from Winter 2020, as it was a transition semester in which the university switched from in-person to remote instruction, and from Fall 2021, as it was a transition semester from remote to in-person instruction. This leads to one Fall and one Winter semester before and during COVID,[7] and excludes what was a largely in-person semester during the pandemic. Thus, the two semesters prior to COVID-19 (Winter 2019 and Fall 2019) represent in-person instruction while the two semesters during COVID-19 (Fall 2020 and Winter 2021) represent largely remote instruction.

We use LEEQ (Section 3.2) to calculate the percentage of words related to each of our teaching qualities for each individual evaluation. Then, we compute the mean values at a per-instructor level both before and during the pandemic. Providing an average per-instructor ensures that instructors who receive more evaluations do not have an outsized effect on the results.

4.1 Differences in Evaluations Before and During COVID-19

We show the results of our analysis in Table 5 and Figure 3; we show histograms of percentage of words in the lexicon for each dimension. This allows us to visualize changes in the data beyond shifts in the mean. As some of the teaching qualities are infrequently mentioned in evaluations, it is not uncommon for averages at an instructor level to be zero. Therefore, as the data is not

[7] We believe that some bias could be introduced by including a higher proportion of Fall semester data, in which more students are adjusting to the university.

Dimension	Pre-COVID	During COVID	Pre-COVID		During COVID	
			Female	Male	Female	Male
High quality	**4.38%**	**4.90%**	4.67%	4.30%	5.14%	4.82%
Helpful	**1.67%**	**1.94%**	1.99%	1.58%	2.21%	1.85%
Clear	**1.01%**	**0.82%**	1.11%	0.98%	0.77%	0.84%
Easy	0.17%	0.14%	0.21%	0.17%	0.14%	0.14%
Low quality	**0.38%**	**0.32%**	0.27%	0.41%	0.26%	0.34%
Unhelpful	0.04%	0.05%	0.02%	0.04%	0.03%	0.06%
Unclear	**0.04%**	**0.02%**	0.02%	0.04%	0.02%	0.02%
Difficult	0.46%	0.36%	0.46%	0.46%	0.33%	0.37%

Table 5: Means pre-COVID and during COVID, both overall and for Female and Male instructors. Significance is computed using the Mann-Whitney U test; **Bold** indicates $p < 0.05$ with FDR adjustment for multiple comparisons (p-values are corrected separately for the overall experiments and the experiments considering instructor sex).

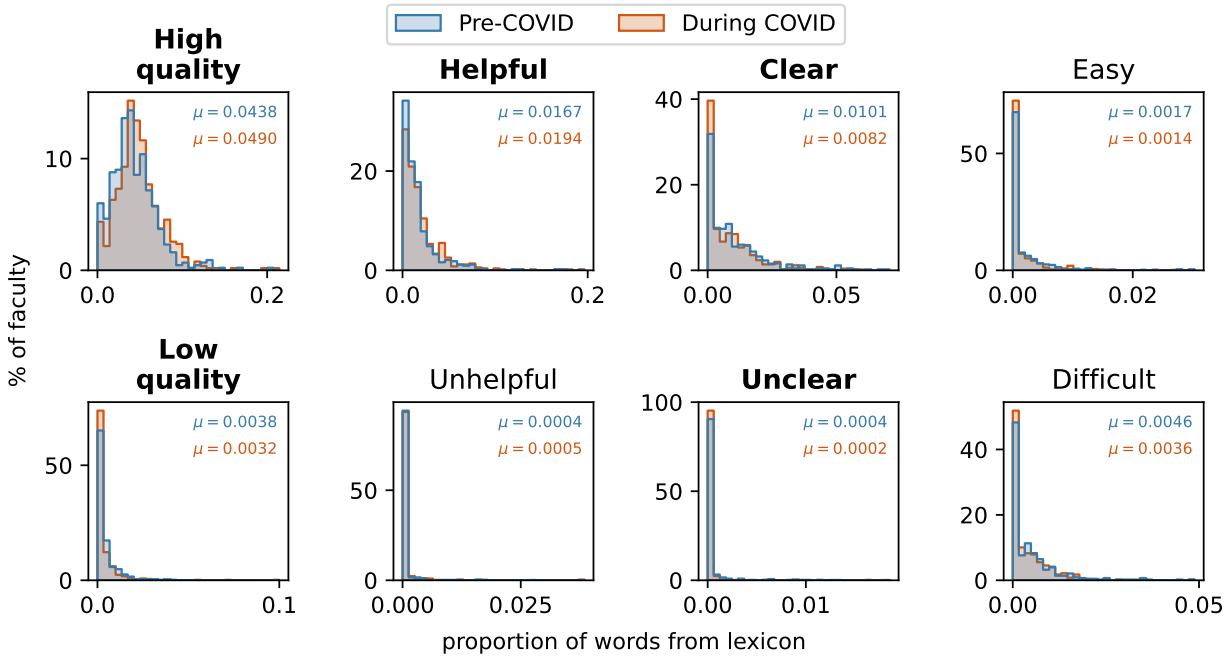

Figure 3: Comparison of evaluations before and during the COVID-19 pandemic using our lexicons. **Bold** indicates $p < 0.05$ with FDR correction for multiple comparisons.

normally distributed, we use the Mann-Whitney U test [28][8] to determine whether there is a statistically significant difference between the percentage of words in each lexical dimension in the evaluations before and during the COVID-19 pandemic. We indicate results that are significant ($p < 0.05$) following an adjustment for false discovery rate (FDR) [24] with $\alpha = 0.05$.

One of the more interesting shifts that we observe is an overall positive shift in the number of

[8]We use this test because the data is not normally distributed.

words that reflect high quality teaching during the COVID-19 pandemic (**high quality** (μ pre-COVID $= 4.38\%$, during = 4.90%) and **low quality** (μ pre-COVID $= 0.38\%$, during = 0.32%)). While we would have expected the shift to remote instruction to be hinder teaching quality, these results indicate that it did the opposite. Some students mention instructor's efforts to set up their course in a way that is conducive to online learning, for instance "the lecture format with modules and discussions was great for online learning." However, other evaluations indicate that students were more lenient when evaluating instructors, given that they know instructors had to put in significant effort to provide a good online experience. This is reflected in comments such as "very good course for being online," indicating that while the student was happy with the instruction, it was only good in the context of an online course, which the comment implies is generally a negative. Another student similarly stated "thank you for doing your best to make this class as enjoyable as possible in this format," indicating that while the format was not ideal, they felt that the instructor did their best. Even in evaluations with **low quality** words such as "the instruction was a bit vague at times, but that is understandable due to the online format," students excuse poor communication due to the format of the course.

Students mention the helpfulness of instructors more in their evaluations during the pandemic (μ pre-COVID $= 1.67\%$, during = 1.94%). One student stated "Professor $PERSON$ was very helpful and understanding of our situation this semester, and has made very clear that he cares about the wellbeing of his students, which I'm very grateful for." Overall, the increase in words in the **helpful** dimension may indicate that instructors were willing to provide some flexibility and grace to students in the face of an unprecedented event, as was shown in prior work [29].

We also see a shift in clarity, with fewer **clear** (μ pre-COVID $= 1.01\%$, during = 0.82%) and fewer **unclear** (μ pre-COVID $= 0.04\%$, during = 0.02%) words during the pandemic. It is interesting to see these lexical dimensions move in the same direction, rather than the opposite direction as we see for **high quality** and **low quality**. To some extent, it may reflect a lack of direct interaction between students and instructors in an asynchronous setting.

The lack of significant changes in either the **easy** or **difficult** dimensions suggests that students did not perceive a major shift in course *content*, only a shift in *delivery*. Overall, the results are encouraging about the quality of remote instruction, although individual evaluations indicate that there is an extent to which *both* students (in their evaluations) and instructors (in their course policies) are being sympathetic *because* of the pandemic itself. Therefore, they likely do not solely reflect the difference between student's perceptions of in-person and online instruction.

4.2 Differences in Evaluations by Instructor Sex

In addition to comparing evaluations prior the pandemic to those during the pandemic, we determine whether there are differences in the percentage of words from each lexical dimension based on the sex of the instructor.[9] We believe that such differences could indicate some bias based on the sex of the instructor. We investigate whether such bias exists, and whether it *changes* during the COVID-19 pandemic. Therefore, we run the analysis on both evaluations before and during the pandemic, and present the results in Table 5 (histograms are available in Appendix B). We use the Mann-Whitney U test to test for significance, and the adjustment for FDR.

[9]This was self-reported and only available as a binary variable, i.e. Male versus Female.

We find no statistically significant differences between male and female instructors either prior to or during the COVID-19 pandemic. This is surprising given evidence of gender bias in teaching evaluations from prior studies (Section 2.2). The dimensions with the most notable and consistent difference, although not statistically significant, are **high quality** (pre-COVID μ M $= 4.30\%$, F = 4.67%, during COVID M $= 4.82\%$, F = 5.14%) and **helpful** (pre-COVID μ M $= 1.58\%$, F = 1.99%, during COVID M $= 1.85\%$, F = 2.21%). This is interesting because it does not reflect a typical definition of bias, which would include *negative* sentiment directed towards a minority group (in this case, women in engineering). We find the difference in the **helpful** dimension to be particularly interesting, as it is reflective of gender differences found in prior work on perceptions of excellent instructors [30]. This work showed that there was more negativity towards female instructors who did not fit gendered expectations. Indeed, while some comments reflecting the helpfulness of female instructors are undoubtedly positive, e.g., "*PERSON* was very helpful when I talked to her in office hours about things I was confused on", some of the words in the **helpful** dimension were clearly used in gendered ways, such as "she was super sweet and very helpful in understanding the labs."[10] The differences between our findings and prior work may be connected a number of factors, including but not limited to the analysis framework, specific characteristics of the university, and changes in perceptions of minority instructors over time. We hope to further examine these differences in future work.

5 Limitations

There are a number of known limitations of dictionary-based lexical analysis. Relying exclusively on word counts ignores additional context in the text. For example, consider the following evaluation: "I thought the class was taught exceptionally well. *PERSON* explained topics so clearly it felt like a friend explaining confusing topics to me in a way they knew I'd understand." The sentiment of evaluation is clearly positive, and the student praises the clarity of the instructor's teaching. However, due to the presence of the word "confusing", the evaluation is given a non-zero score for **low quality**.

This challenge is magnified in the case of negations. For example, "sometimes, the explanations of concepts were a *not* clear, and we were never able to finish the lectures" indicates lack of clarity, but the presence of "clear" would increase the score for that dimension. This is less likely to be a problem when the dimension itself represents a negative aspect of teaching, as double negatives such as "not unclear" appear infrequently.

Finally, the somewhat low coverage of our lexicon shown in Section 3.2 means that a large number of lexicon scores are 0, even when we average across all evaluations of an instructor. A more broad-coverage approach might use qualitative methods such as open-ended coding [31], to make the most out of the data set, although this poses its own issues with scalability. Still, we believe that our lexicon is more useful for understanding teaching evaluations than more general lexicons like LIWC [26], which does not capture teaching-specific language.

While these problems likely have some effect on our analysis, the direction of correlations between lexical dimensions and the "Excellent Teacher" ratings (see Table 3.3) indicate that they generally capture teaching quality as expected. Notably, while the lexicon is useful for studying

[10] While female instructors make up approximately 25% of the dataset, 8/10 evaluations containing "sweet", "sweetest", and "sweetheart" pertained to female instructors.

aggregate patterns in the data, there may be too much noise to reliably apply it to understand individual evaluations (e.g., whether a particular evaluation states that the instructor was more "helpful" than other evaluations).

6 Conclusion

In this study, we developed a lexicon (LEEQ) with eight dimensions related to teaching ability, using a combination of NLP methods and manual filtering. We used LEEQ to compute the percentage of words related to eight teaching qualities for instructors in a secondary dataset from a large U.S. university. We confirmed the validity of LEEQ by examining each dimension's correlation with instructor's ratings, which tended to be high for positive teaching qualities and low for negative teaching qualities. Furthermore, we found that the percentage of words in pairs of dimensions representing positive teaching qualities tends to be highly correlated to the percentage of words representing other positive teaching qualities (and vice-versa for negative qualities). Finally, we compare our lexicon's correlations with ratings given by existing sentiment analysis methods, and find that they outperform simple approaches like LIWC and are competitive overall, while remaining more closely related to the teaching domain than existing lexicons.

In a case study, we examined changes in evaluations during the pandemic. We found a surprising increase in words from the **high quality** dimension and a decrease in words from the **low quality** dimension. Upon closer examination of the evaluations, we found that some students contextualize the use of positive words with statements like "for an online course," which led us to believe that students might be more lenient in their evaluation of instructors during the pandemic. On the other hand, student's increased use of words from the **helpful** dimension indicates that instructors went out of their way to accommodate students during the shift to online instruction. Finally, we observed words from the **clear** and **unclear** dimensions decrease during the pandemic. We did not find significant differences between male and female instructors; however, we did find small shifts in language use that suggest that some gendered language is present in the evaluations.

The lexicon developed in this study can inform future work that conducts large-scale analysis of student comments. We designed our lexicon to have high precision rather than high recall, which is reflected in the relatively low overall frequencies (Table 2). To this last point, we recommend that future researchers build on this lexicon and cast a wider "net" to capture students' perceptions of teacher performance. By improving our methods for text analysis, we can improve our "lens" for identifying possible forms of bias against instructors and changes over time, and provide support for instructors who may need it.

Acknowledgements

This work was partially supported by the Templeton Foundation (#62256). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Templeton Foundation.

References

[1] Susan Basow, Stephanie Codos, and Julie Martin. The effects of professors' race and gender on student evaluations and performance. *College student journal*, 47(2):352–363, 2013.

[2] Anne Boring. Gender biases in student evaluations of teaching. *Journal of public economics*, 145:27–41, 2017.

[3] Monica Ciolacu, Ali Fallah Tehrani, Rick Beer, and Heribert Popp. Education 4.0—fostering student's performance with machine learning methods. In *2017 IEEE 23rd international symposium for design and technology in electronic packaging (SIITME)*, pages 438–443. IEEE, 2017.

[4] Bob Heller, Mike Proctor, Dean Mah, Lisa Jewell, and Bill Cheung. Freudbot: An investigation of chatbot technology in distance education. In *EdMedia+ Innovate Learning*, pages 3913–3918. Association for the Advancement of Computing in Education (AACE), 2005.

[5] Ho Thao Hien, Pham-Nguyen Cuong, Le Nguyen Hoai Nam, Ho Le Thi Kim Nhung, and Le Dinh Thang. Intelligent assistants in higher-education environments: the fit-ebot, a chatbot for administrative and learning support. In *Proceedings of the ninth international symposium on information and communication technology*, pages 69–76, 2018.

[6] Norman Meuschke and Bela Gipp. State-of-the-art in detecting academic plagiarism. *International Journal for Educational Integrity*, 9(1), 2013.

[7] Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. Wronging a right: Generating better errors to improve grammatical error detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4977–4983, 2018.

[8] Zackary Dunivin, Lindsay Zadunayski, Ujjwal Baskota, Katie Siek, Jennifer Mankoff, et al. Gender, soft skills, and patient experience in online physician reviews: a large-scale text analysis. *Journal of medical Internet research*, 22(7):e14455, 2020.

[9] Angie Waller and Kyle Gorman. Detecting objectifying language in online professor reviews. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 171–180, 2020.

[10] Nathan E Gonyea and Joseph M Gangi. Reining in student comments: a model for categorising and studying online student comments. *Assessment & Evaluation in Higher Education*, 37(1):45–55, 2012.

[11] Ellyn Kaschak. Sex bias in student evaluations of college professors. *Psychology of Women Quarterly*, 2(3):235–243, 1978.

[12] Kristina MW Mitchell and Jonathan Martin. Gender bias in student evaluations. *PS: Political Science & Politics*, 51(3):648–652, 2018.

[13] Nikolas Gordon and Omar Alam. The role of race and gender in teaching evaluation of computer science professors: A large scale analysis on ratemyprofessor data. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pages 980–986, 2021.

[14] Andrew S Rosen. Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of ratemyprofessors. com data. *Assessment & Evaluation in Higher Education*, 43(1):31–44, 2018.

[15] Yanan Fan, Laura J Shepherd, Eve Slavich, David Waters, M Stone, Rachel Abel, and Emma L Johnston. Gender and cultural bias in student evaluations: Why representation matters. *PloS one*, 14(2):e0209749, 2019.

[16] Kenneth M Cramer and Louise R Alexitch. Student Evaluations of College Professors: Identifying Sources of Bias. *Canadian Journal of Higher Education*, 30(2):143–64, 2000.

[17] Beatrice Tucker. Student evaluation surveys: Anonymous comments that offend or are unprofessional. *Higher Education*, 68(3):347–358, 2014.

[18] Tamara Baldwin and Nancy Blattner. Guarding against potential bias in student evaluations: What every faculty member needs to know. *College Teaching*, 51(1):27–32, 2003.

[19] Wendy M Williams and Stephen J Ceci. "how'm i doing?" problems with student ratings of instructors and courses. *Change: the magazine of higher learning*, 29(5):12–23, 1997.

[20] M. Azab, R. Mihalcea, and J. Abernethy. Analysing ratemyprofessors evaluations across institutions, disciplines, and cultures: The tell-tale signs of a good professor. In *Proceedings of the 8th International Conference on Social Informatics (SocInfo 2016)*, Bellevue, WA, 2016.

[21] Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. Deconfounded lexicon induction for interpretable social science. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1615–1625, 2018.

[22] William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, 2016.

[23] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[24] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL `http://www.jstor.org/stable/2346101`.

[25] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014. doi: 10.1609/icwsm.v8i1.14550. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/14550`.

[26] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015. URL `https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf`.

[27] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 54–59, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-4010. URL https://aclanthology.org/N19-4010.

[28] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60, 1947. doi: 10.1214/aoms/1177730491. URL https://doi.org/10.1214/aoms/1177730491.

[29] Susannah C. Davis, Yan Chen, Vanessa Svihla, Madalyn Wilson-Fetrow, Pil Kang, Abhaya K. Datye, Eva Chi, and Sang M. Han. Pandemic pivots show sustained faculty change. In *2021 ASEE Virtual Annual Conference Content Access*, Virtual Conference, July 2021. ASEE Conferences. https://peer.asee.org/37557.

[30] Joey Sprague and Kelley Massoni. Student evaluations and gendered expectations: What we can't count can hurt us. *Sex Roles*, 53(11):779–793, Dec 2005. ISSN 1573-2762. doi: 10.1007/s11199-005-8292-4. URL https://doi.org/10.1007/s11199-005-8292-4.

[31] Elizabeth Davison and Jammie Price. How do we rate? an evaluation of online student evaluations. *Assessment & Evaluation in Higher Education*, 34(1):51–65, 2009.

A Rate My Professors Rating Form

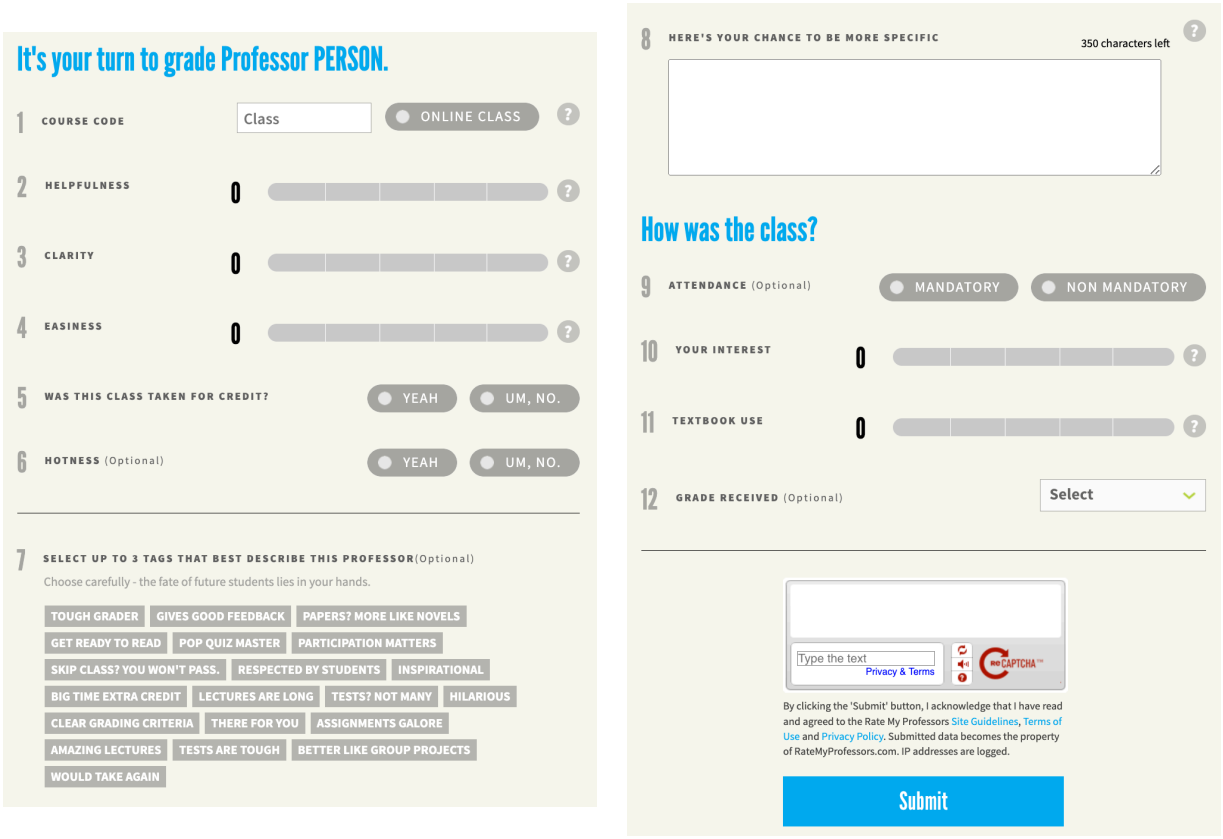

Figure 4: Screenshot of the rating form from the Rate My Professors website. The historical 2014 version of the form was retrieved from archive.org, and the instructor's name is redacted for privacy.

B Plot of comparisons by instructor sex

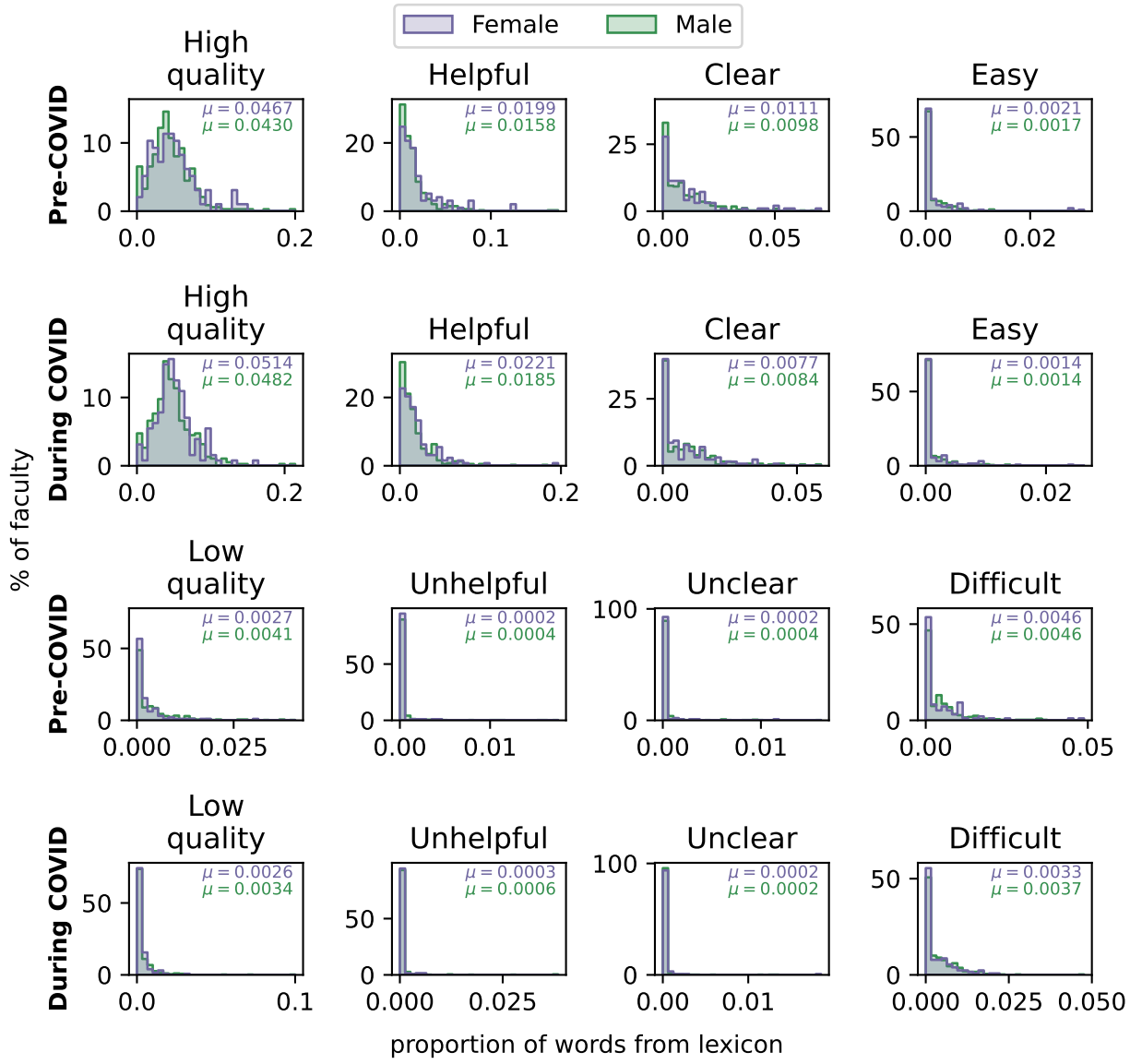

Figure 5: Comparison of evaluations split by instructor Sex before and during the COVID-19 pandemic using our lexicons. **Bold** indicates $p < 0.05$ with FDR correction for multiple comparisons.