

Lexical Measurement of Teaching Qualities

Laura Biester, Ian Stewart, **Laura Hirshfield**, Rada Mihalcea, Sara Pozzi

June 28, 2023
ASEE



Motivation

- Course evaluations are often biased, both in
 - Ratings: male instructors are rated higher than female instructors¹
 - Language: abusive language is used to describe female instructors²
- In addition to being intrinsically problematic, this bias hurts minority instructors due to reliance on course evaluations during tenure and promotion^{3,4}
- **Overall project research questions:**
 - How did evaluations change as a result of remote learning?
 - Was bias amplified or subdued?

[1] Boring (2017), Gender biases in student evaluations of teaching

[2] Tucker (2014), Student evaluation surveys: Anonymous comments that offend or are unprofessional

[3] Baldwin and Blattner (2003), Guarding against potential bias in student evaluations: What every faculty member needs to know

[4] Williams and Ceci (1997), "How'm I Doing?" Problems with Student Ratings of Instructors and Courses

Motivation

- Text in course evaluations may reveal differences in how students **describe** instructors that are not present in Likert-scale data
- Sentiment analysis captures only **one dimension** of teaching (quality)
 - Differences in **specific dimensions** may be more informative, and may differ between time periods and demographic groups

We created a lexicon with 8 dimensions of teaching quality:
LEEQ (Lexicon for Evaluation of Education Quality)

LEEQ Creation: Our Dataset

- Evaluations of **1,075 instructors** before and during the COVID-19 pandemic
 - Winter 2019 – Fall 2021
- Includes both **ratings** (Likert scale) and **free-response question(s)** about teaching quality

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

Note: limitation of registrar data recording “sex” and not gender

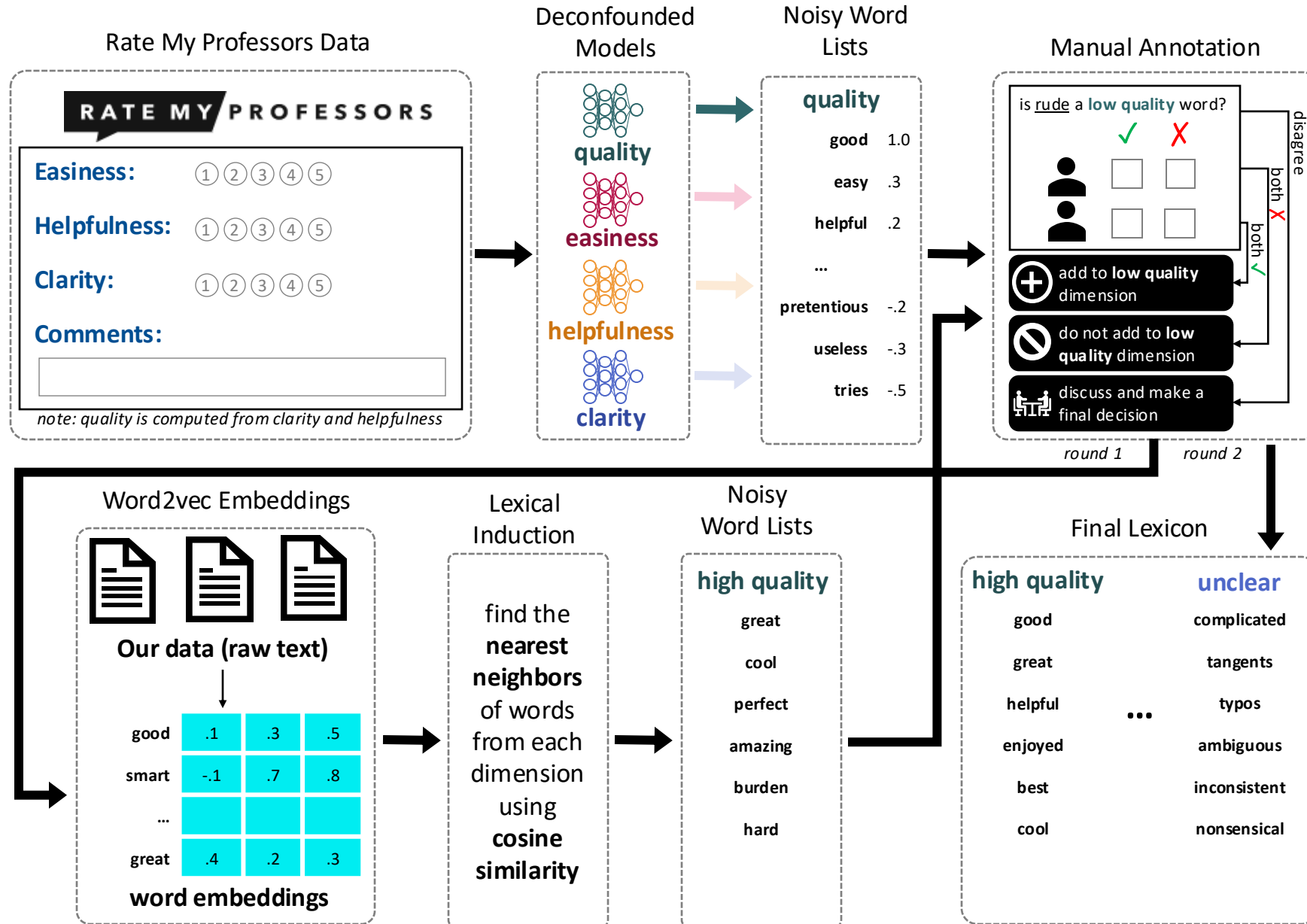
LEEQ Creation: Initial Dataset

- Based on 863,857 Rate My Professors reviews covering 31 universities
- Each review includes
 - Numerical ratings for helpfulness, clarity, easiness
 - Free-text input

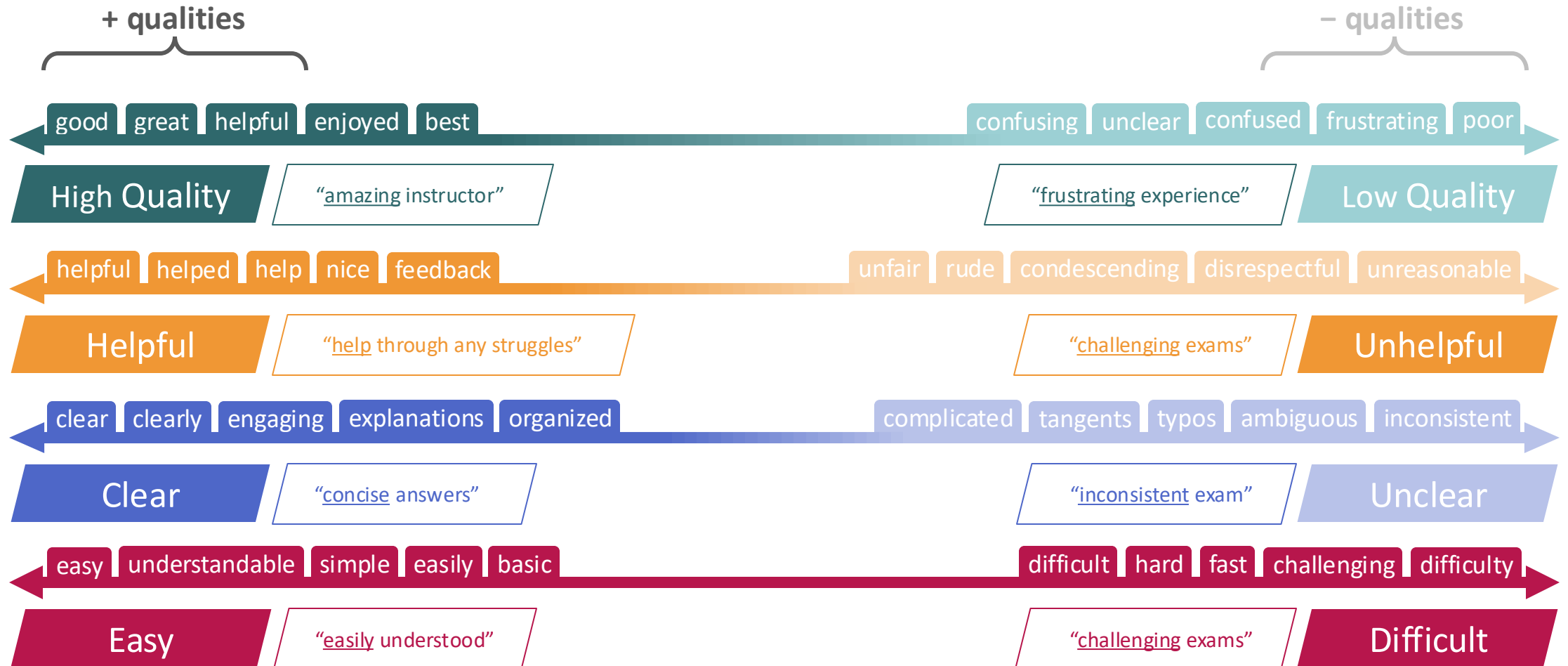
The screenshot displays the 'Rate My Professors' review form, which is used for creating the LEEQ dataset. The form is divided into several sections, each with a numbered heading and a question. The sections are:

- 1 COURSE CODE**: A text input field labeled 'Class' and a radio button for 'ONLINE CLASS'.
- 2 HELPFULNESS**: A numerical rating from 0 to 5, with a red box highlighting the rating area.
- 3 CLARITY**: A numerical rating from 0 to 5, with a red box highlighting the rating area.
- 4 EASINESS**: A numerical rating from 0 to 5, with a red box highlighting the rating area.
- 5 WAS THIS CLASS TAKEN FOR CREDIT?**: Radio buttons for 'YEAH' and 'UM, NO.'.
- 6 HOTNESS (Optional)**: Radio buttons for 'YEAH' and 'UM, NO.'.
- 7 SELECT UP TO 3 TAGS THAT BEST DESCRIBE THIS PROFESSOR (Optional)**: A list of tags including 'TOUGH GRADER', 'GIVES GOOD FEEDBACK', 'PAPERS? MORE LIKE NOVELS', 'GET READY TO READ', 'POP QUIZ MASTER', 'PARTICIPATION MATTERS', 'SKIP CLASS? YOU WON'T PASS.', 'RESPECTED BY STUDENTS', 'INSPIRATIONAL', 'BIG TIME EXTRA CREDIT', 'LECTURES ARE LONG', 'TESTS? NOT MANY', 'HILARIOUS', 'CLEAR GRADING CRITERIA', 'THERE FOR YOU', 'ASSIGNMENTS GALORE', 'AMAZING LECTURES', 'TESTS ARE TOUGH', 'BETTER LIKE GROUP PROJECTS', and 'WOULD TAKE AGAIN'.
- 8 HERE'S YOUR CHANCE TO BE MORE SPECIFIC**: A text input field with a character count of 350 characters left.
- 9 ATTENDANCE (Optional)**: Radio buttons for 'MANDATORY' and 'NON MANDATORY'.
- 10 YOUR INTEREST**: A numerical rating from 0 to 5.
- 11 TEXTBOOK USE**: A numerical rating from 0 to 5.
- 12 GRADE RECEIVED (Optional)**: A dropdown menu labeled 'Select'.

At the bottom of the form, there is a text input field for a free-text review, a reCAPTCHA verification, and a 'Submit' button. A disclaimer at the bottom states: 'By clicking the "Submit" button, I acknowledge that I have read and agreed to the Rate My Professors Site Guidelines, Terms of Use and Privacy Policy. Submitted data becomes the property of RateMyProfessors.com. IP addresses are logged.'

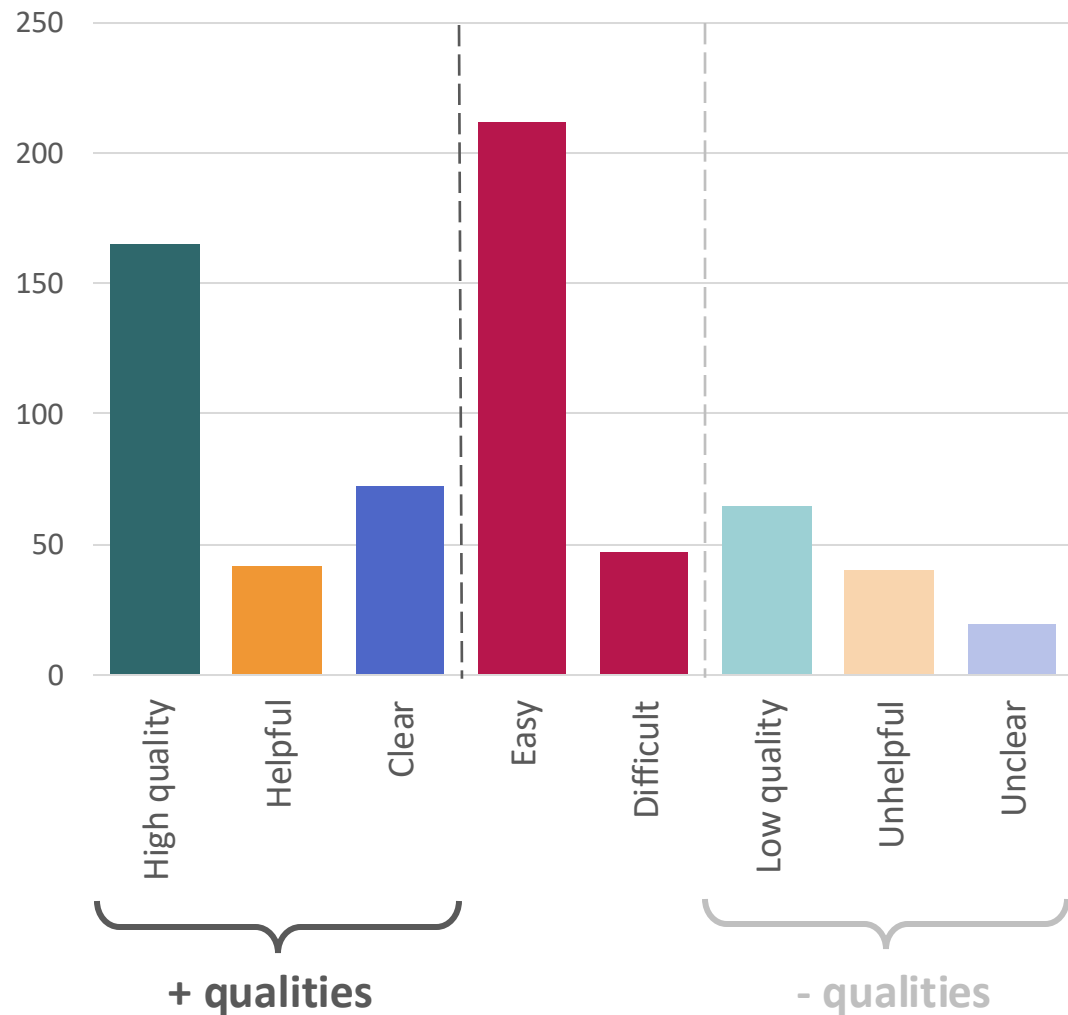


Examples

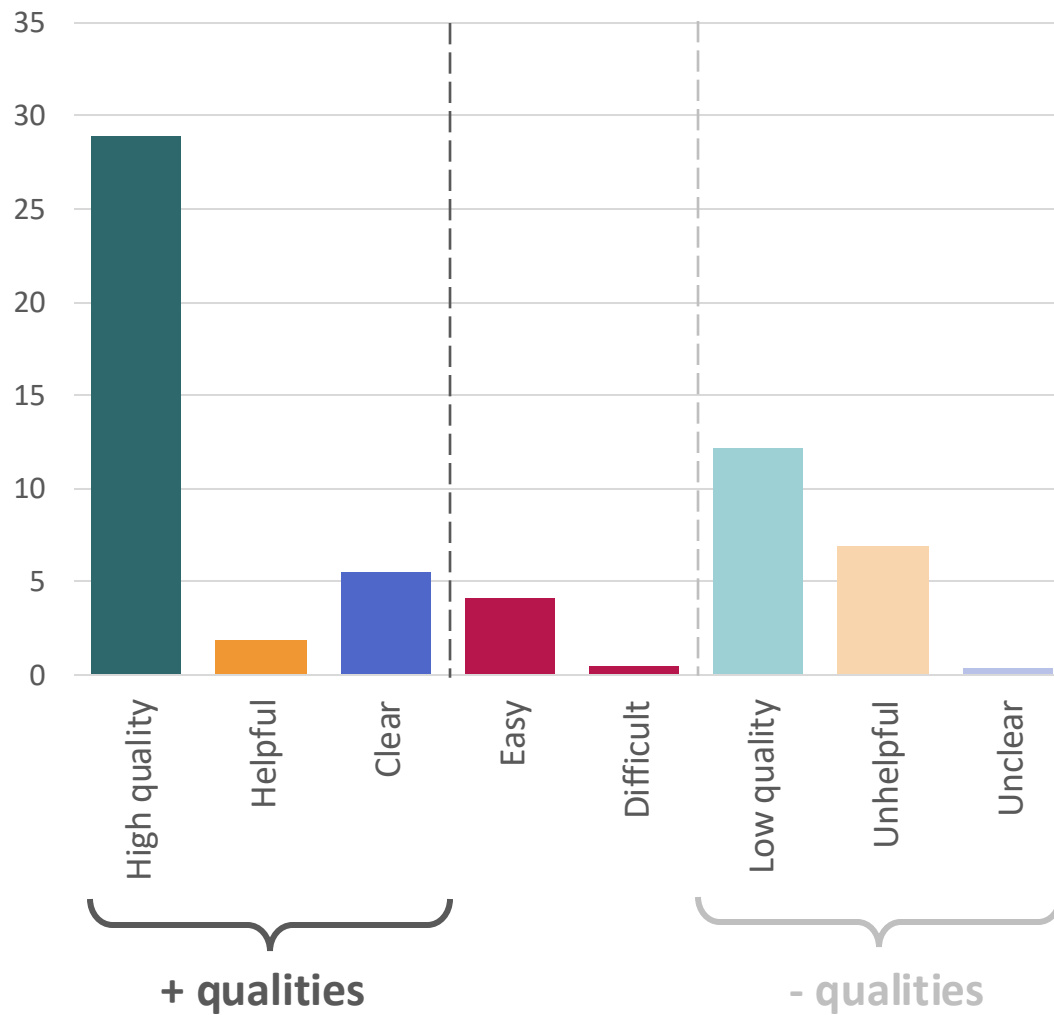


Uniqueness & Frequency

Unique Words



Freq (per 1K words)



Validation

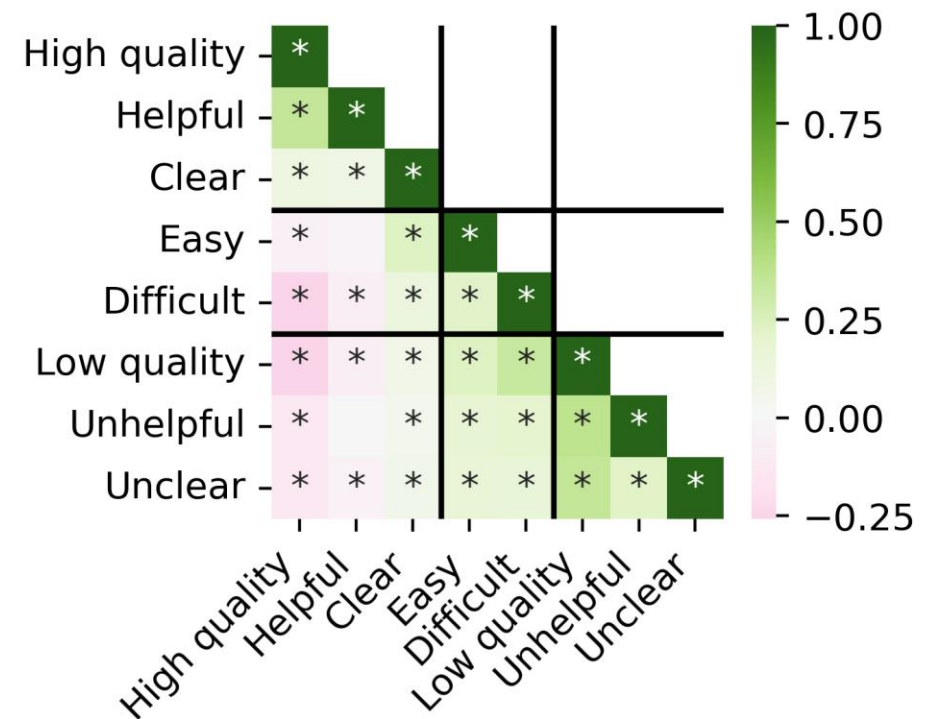
External Validation: Numerical Ratings

Dimension	Correlation
High quality	0.3931*
Helpful	0.0822*
Clear	0.0086
Easy	-0.0685
Difficult	-0.2754*
Low quality	-0.4281*
Unhelpful	-0.2292*
Unclear	-0.2184*

Correlations between % of words from each lexicon and instructor's mean numerical ratings

*** denotes $p < 0.05$**

Internal Validation: With Each Other



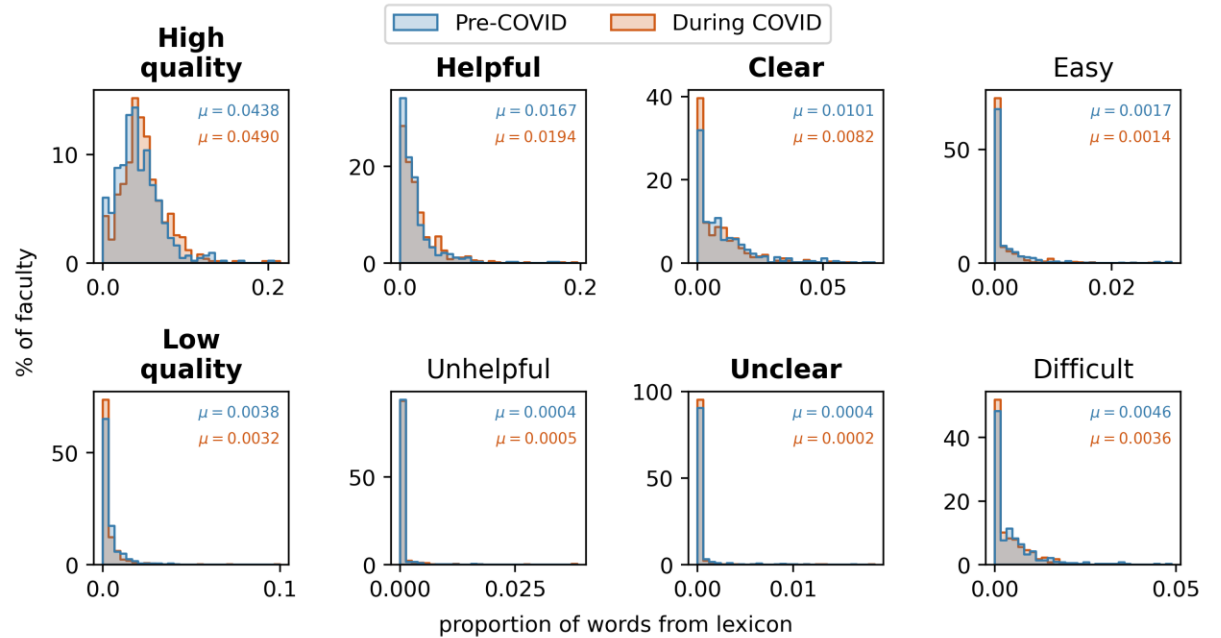
Correlations between % of words from pairs of lexicons

*** denotes $p < 0.05$**

Case Study

- What differences occur in free-text course evaluations between **in-person and remote/hybrid learning**?
 - We use 4 semesters, dropping Winter 2020 and Fall 2021, as these semesters encompassed transitions between teaching formats
- What differences arise between **instructors of different identities** *in each period*?
 - We consider instructor sex for our case study

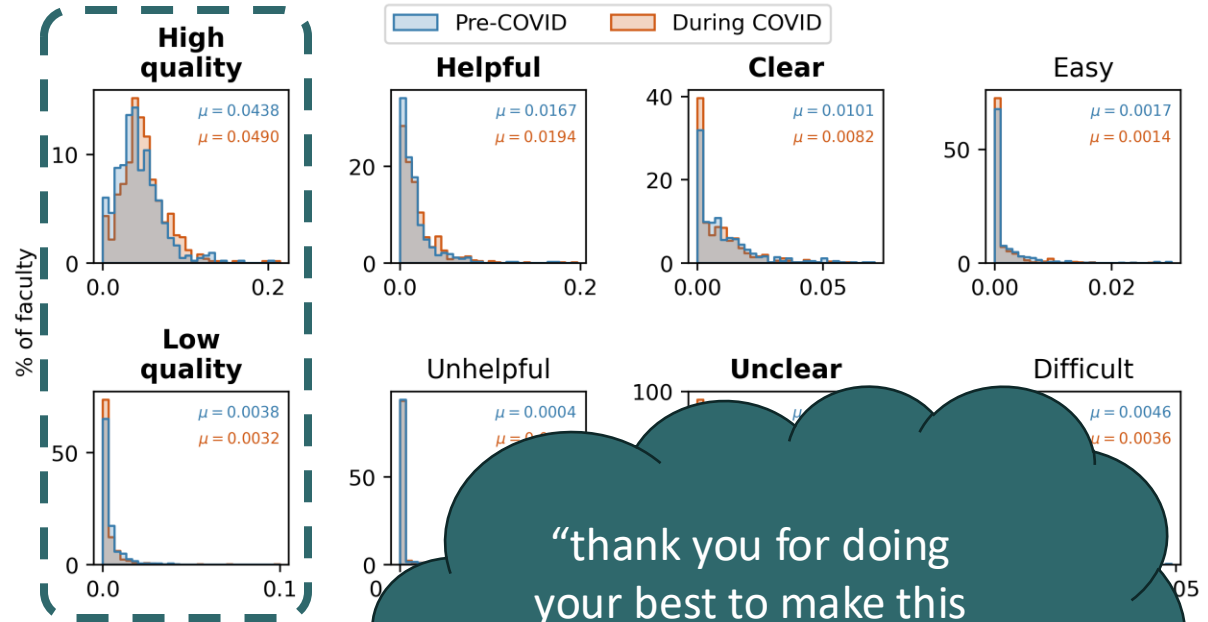
Differences between in-person and remote learning



bold $\rightarrow p < 0.05$ (Mann-Whitney U test)

Differences between in-person and remote learning

- **Quality** doesn't suffer during COVID
 - high quality $\uparrow$ and low quality $\downarrow$

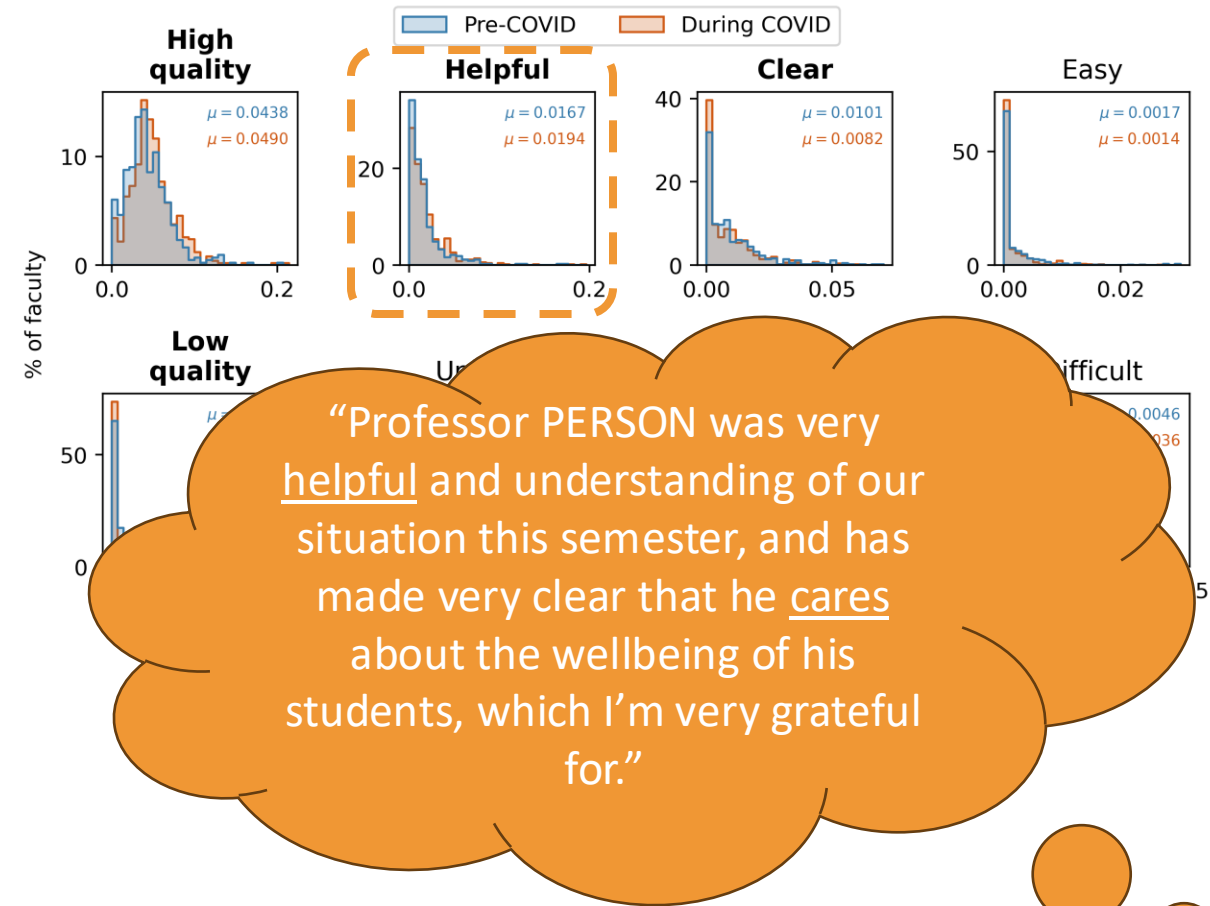


bold $\rightarrow$ p

"thank you for doing your best to make this class as enjoyable as possible in this format"

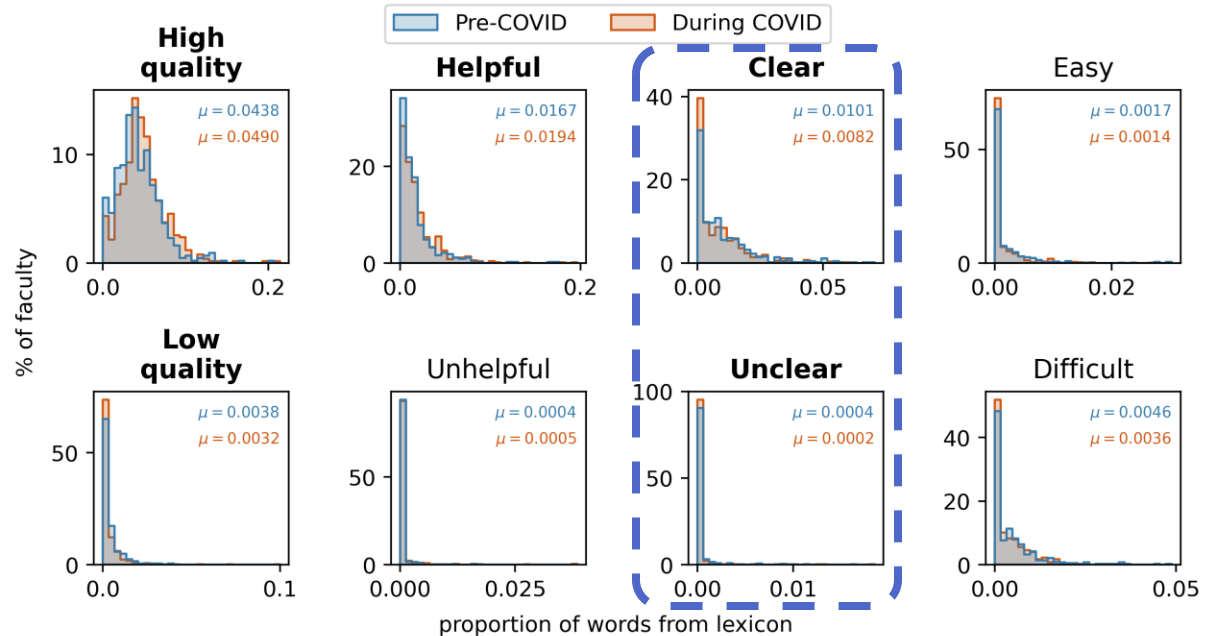
Differences between in-person and remote learning

- **Quality** doesn't suffer during COVID
 - high quality $\uparrow$ and low quality $\downarrow$
- Students use more **helpful** words to describe instructors during the pandemic



Differences between in-person and remote learning

- **Quality** doesn't suffer during COVID
 - high quality $\uparrow$ and low quality $\downarrow$
- Students use more **helpful** words to describe instructors during the pandemic
- Fewer words related to **clarity**, either positive or negative

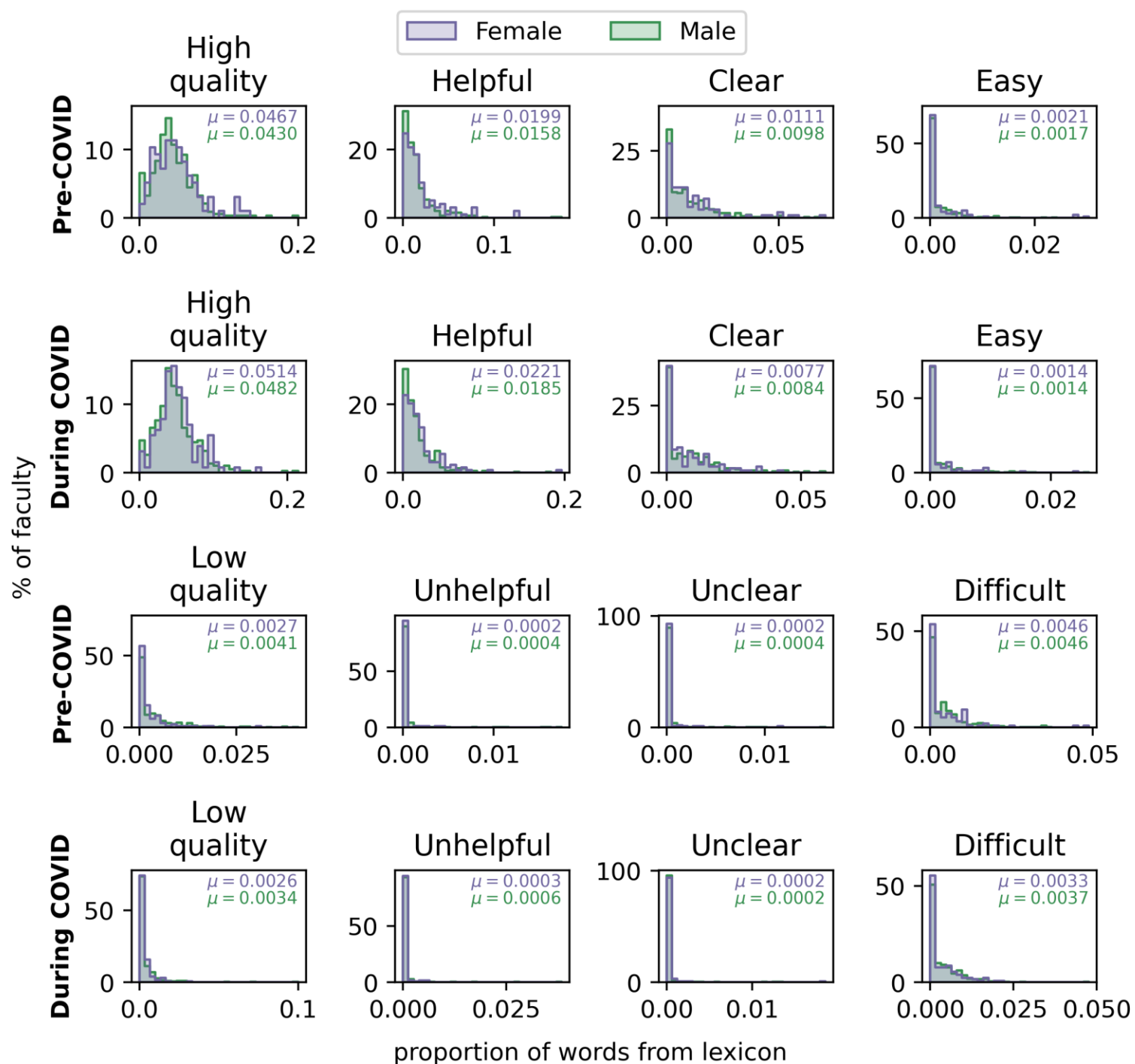


bold $\rightarrow p < 0.05$ (Mann-Whitney U test)

COVID + Gender: Results

- There are no significant differences before or after COVID-19
- Some gendered language still appears in the dataset, e.g., “she was super **sweet** and very helpful in understanding the labs.”

80% of mentions of “sweet” involve female instructors, even though only 25% of instructors are female



Conclusions

Future work:

- Analysis of additional demographic attributes
- Analysis of numerical ratings in the CCE dataset

Our lexicon is freely available at
github.com/MichiganNLP/LEEQLexicon



Lexicon