Evaluations as Research Tools: Gender Differences in Academic Self-Perception and Care Work in Undergraduate Course Reviews

David Lang, Youjie Chen, Andreas Paepcke, Mitchell L. Stevens
Stanford University
Stanford, CA
{dnlang86, minachen, paepcke, stevens4}@stanford.edu

ABSTRACT

Student course reviews are rarely considered as research instruments, yet their ubiquity makes them promising tools for education data science. To illustrate this potential, we use a corpus of student reviews to observe gender differences in how students appraise their own learning and in the advice they give to to future students. We find systematic differences in who submits course reviews, with female and academically high-achieving students more likely to submit. Among submitters, we find (a) females understate their achievement of learning goals relative to males earning the same grades; (b) females offer lengthier written advice to future students than males; (c) advice written by females exhibits more positive tone, even after accounting for grades and course selections.

Keywords

care work; course evaluations; gender; higher education; topic models; survey design [1]

1. INTRODUCTION

Student course reviews are a controversial subject in academia. While considerable work has addressed problems of validity and bias in the use of these instruments for assessing instructors and instruction [25] [21], few have recognized reviews as potentially useful research tools for education data science.

Several features of course reviews make them potentially attractive for researchers. First, reviews are ubiquitous features of teaching and learning in US higher education. Because they are so commonly solicited and so frequently submitted, the data yielded from reviews represents a very wide swath of student populations. Second, reviews are routinely submitted through online platforms and carried out by administrative units supported on hard budget lines, bringing the marginal cost of acquiring research data through reviews close to zero. Third, it is technically simple to link information obtained through reviews with institutional data describing those who submit them. Thus while course reviews may be problematic means of assessing the quality of instructors and instruction – a matter on which we make no comment here– we believe these data hold substantial promise for education data science.

To illustrate this promise, we leverage a corpus of 11,255 student reviews submitted by undergraduate students enrolled in Computer Science (CS) classes at a private research university during the 2015-16 and 2016-17 academic years. Because each review is linked with the academic transcript and self-reported gender of its submitter, we are able to observe variation in submissions by an important aspect of student identity and documented academic accomplishment.

While a variety of student characteristics of are of interest to education data scientists, we focus on students' gender for reasons both practical and theoretical. For privacy purposes, our case university has currently granted researcher access to only a few variables describing review submitters; we utilize those available data here. Yet we also have two theoretical motivations for focusing on gender. First, borrowing from social psychology, we recognize that women tend to under-estimate their own abilities, while men to to over-estimate, conditional on measured accomplishment [7]. Second, borrowing from feminist social science, we posit that submission of a course review is a form of care work – a voluntary investment in the well-being of others – and thus implicated differently in feminine and masculine gender roles and identities [9]. Our findings comport with the contours of these larger literatures in ways that are both important in their own right, and instructive for any future deployments of course reviews for education data science.

We pursue three sets of analyses below. In the first set, we observe variation in rates of review submission by gender and earned grades. These analyses illustrate how researchers might test the representativeness of corpora of reviews. Second, we observe how submitters respond to multiple close-ended review prompts targeting self-assessments of learning, but that are phrased differently. These analyses illustrate how review design may interact with student characteristics to produce patterned variation in reported learning progress. Third, we conduct computational text analyses of submissions to an open-ended review prompt. These analyses illustrate how qualitative reviews can be efficiently leveraged for scientific insight.

2. RELATED WORK

2.1 Course Reviews

Research on course reviews typically has focused on questions of their value as an instruments for evaluating the quality of instruction. Analyses conducted at scale typically focus on whether measures of learning are correlated with instructional quality [27].

One potential concern about course reviews from the pscyhometrics liaterature is the potential for differential item function, a phenomenon in which respondents of equal ability will exhibit different responses to a given survey item or question. Studies of differential item function in course reviews have focused on the quantitative difficulty of a class, or characteristics of the instructor [17] [8]. Relatively little work has focused on how student characteristics may be associated with differential item function on course reviews.

If findings from copious research on product reviews translate to academic course reviews, we would expect that students with high-valence opinions about a course are more likely to respond, resulting in a bimodal or *j-shaped* distribution [12]. In practice, these findings may not translate. There are often other incentives for filling out reviews, for example, giving students earlier access to their final grades as an incentive to respond. Those who submit reviews may not be representative of the larger population of students who enrolled in a particular course, or of the overall campus population. This problem is exacerbated when analysts to not have access of reviewer characteristics such as gender or grades [1]. Past work that has tried to adjust for non-response bias in review has suggested that non-response bias tends to favor positive reviews [11].

There is varying theory around why students may opt to submit reviews. Studies conclude that students are more likely to respond to course evaluations if they are majoring in the subject of the course [1]. Other work suggests that female students are generally more likely to respond to course evaluations than males [23]. However, little work has focused specifically on this topic in Computer Science courses.

Under experimental conditions where researchers manipulated the information content and valence of course reviews, researchers found that these factors had material effects on course enrollment decisions. Students were more likely to enroll in courses if course evaluations had positive valence, particularly if there was a large number of such evaluations [14]. Similar work found that exposure to positive or negative course reviews had modest to large effects on students' expected performance within a course, and their likelihood of recommending the course in the future[13]. These findings are particularly relevant for CS courses as CS courses tend to have relatively enrollments compared to other subjects.

2.2 Text Analysis

There is a burgeoning literature on using computational text analysis and methods to quantify differences in corpora based on characteristics of the author and the text. These techniques have been quickly adopted to educational applications but we have seen relatively few instances of text analysis of course reviews. When text analysis of course evaluations are done, they are typically focused on keyword extraction and on predicting Likert item responses as a function of the text [24].

We group text analyses methods into the following three categories:

2.2.1 Dictionary and rule-based approaches

Dictionary-based approaches characterize the words of documents into groups of predefined categories such as sentiment. The most popular of these dictionaries is the Linguistic Inquiry and Word Count (LIWC) dictionary [18]. In addition to grouping words into 75 distinct categories and themes (e.g. family, power, death, etc), the dictionary generates four psycho-social variables that were validated on college application essays through a rating process. Each of these variables is scored on a 1 to 99 interval where 1 is a complete lack of the construct or and 99 is highly pronounced form of the construct. These constructs are:

1. **Tone**- This is a summary variable describing the emotional quality of the text. A score of 99 reflects a positive tone and a score of 1 reflects a negative tone. A score of 50 represents neutral valence.
2. **Analytic**- This is a measure of how much formal logic is used in the text. A score of 1 indicates little use of formal logic and a score of 99 exhibits statement with a great deal of formal logic.
3. **Authenticity**- This is a measure of the sincerity/honest of a text. A score of 1 indicates insincerity and a score of 99 indicates high sincerity.
4. **Clout**- This is a measure of the text's authority, relative position, and confidence. A score of 1 suggests relatively little authority and a score of 99 suggests high authority.

Researchers can also create their own custom measures of text. We take advantage of this affordance by capturing mentions of instructors' names.

2.2.2 Token-based approaches

Token-based approaches treat every word in a text as input into a model. These approaches often result in the loss of syntactic meaning but are often very effective at classifying documents. Token-based approaches have proven effective at detecting socioeconomic features of authors such as race, gender, and income in college application essays [4]. Other applications have generated algorithms with high predictive validity on classroom observation and evaluation rubrics [15].

2.2.3 Unsupervised approaches

The basic premise behind unsupervised approaches is that texts include multiple topics, and topics comprise words. Using unsupervised methods such as Latent Dirichelet Allocation (LDA) , we can group texts categorically. These same methods have been augmented recently to allow the distribution of topics to co-vary with other relevant metadata, a technique known as *structural topic modeling* [22]. In this case, we can examine the concentration of topics by features such as student gender or grades. This method further allows us to perform statistical inference to see if topic preponderance varies systematically by characteristics of authors.

2.3 Gender Differences in Academic Experiences, Skill Perception and Care Work

Our work has three motivations from prior social-science literature on higher education and gender. The first is that male and female students may have different experiences when taking the same courses. For example, women are less comfortable asking questions and have less confidence in CS courses than their male peers [19]. This "gender confidence gap" grows as students take more advanced courses [3]. Analyses of communal academic resources in CS programs

find substantial differences in how contributions by male and female users are acknowledged *Github* and *StackOverflow* [16] [26]. Consequences of these phenomena may extend beyond college,as women with degrees in STEM fields are less likely than men to enter STEM occupations [5]. While course reviews cannot capture empirical variation in experience per se, they can capture how submitters make sense of those experiences.

Second, gendered differences in skill perception may influence how students report their experiences and learning gains in reviews. While women tend to approach STEM fields with less confidence, men tend to over-estimate their abilities. Experimental work by Correll [7] found that men expressed inflated perceptions of their own skill at completing quantitative tasks compared to women performing at the same level of measured accomplishment. Together these inquiries suggest that course reviews may bear traces of gendered patterns of academic self-perceptions. Our third motivation is the gendered character of care work. Social scientists define care work as work that attends to the well-being of others. It comprises activities and services intended to help other people develop their capabilities and pursue their goals [9]. Care work is consistently associated with femininity and female role expectations, and often is unpaid or poorly compensated [10]. To the extent that submitting course reviews is an act of assistance – to improve classes and to inform future students – it is appropriately theorized as a form of care work. Thus we might expect that female and male students will approach the task of course reviews with different dispositions, such that the number, extensiveness, and content of course evaluations may vary by gender of submitters.

3. RESEARCH QUESTIONS

Our working hypothesis is that course reviews will exhibit gendered patterns of academic experience, self-perceptions and advice-giving. Specifically: (1) reviews from male students will exhibit stronger professed strong learning gains (2) reviews from female students will exhibit characteristics of care work.

We group our analyses into two parts. The first part examines variation by gender and earned grades on review submission rates and on Likert-scale items on course reviews. The second part examines variation in male and female responses to a qualitative review prompt eliciting advice for future students considering the same courses.

3.1 Review Submission Rates and Likert Items

H1: Female students will respond to course evaluations more often than males.

Our care work hypothesis is that female students will be more responsive to institutional requests for reviews. We investigate this hypotheses using an exact binomial-two-sample test. We examine results by gender and grade.

H2: There are systematic differences in response rate by grade. There are many competing theories of how grades might influence response rates to course evaluations. If students have a poor grade, they may be more inclined to view the evaluation as an opportunity to retaliate against the grader. Alternatively, students who receive a low grade may opt to avoid opportunities to reflect on negative experiences. We will investigate this hypothesis utilizing a simple χ^2 test of response-rates by grade.

H3: Female students will understate their achievements relative to their male counterparts. We hypothesize that men are more likely to see course reviews as a form of positive self-reflection and promotion, and that females are more likely reviews as a form of care work. We believe these differences will have stronger valence in items that focus on a students' accomplishments rather than other constructs such as student learning. We will model these analyses as a fixed-effect regression model with the following specification:

$$Y_{ij} = \beta_1 Male_i + Grades_{ij} + \Gamma_j + \varepsilon_{ij} \quad (1)$$

The subscripts i and j correspond to indices for student and course. The Y variable corresponds to our focal outcome variable, in this case, responses to a Likert item. $Male$ corresponds to a student's self-reported indicator variable of whether the student identifies as male and β_1 corresponds to the associated coefficient with this variable. We represent course effects with Γ_i to control for factors like the difficulty of the course or instructional quality. We also control for grades with an additional fixed-effect for each possible grade a student could receive [2]. The error is represented by ε. Errors are clustered at the course level.

3.2 Open-Response Questions

We pay particular interest to open-response items in course reviews. We suspect that such items are may be the most valuable and least explored element of course reviews. As such, we may be able to detect subtle differences in qualitative responses.

3.2.1 Psychosocial variables

H4: Course evaluations written by females will express more positive and sincere sentiment.

Given our care work hypothesis, we believe that female students will express more positive sentiment in open-response items. We use the same analytical strategy as an equation 1 using LIWC's tone variable. Specifically, we examine gender differences in these psycho-social variables after controlling for variation that can can be attributed to the course, or to student grade. We report outcomes in standardized effect sizes to facilitate interpretability.

We also hypothesize that a corollary to the care work hypothesis is that female students will use more "I" statements and tentative language. This tendency would manifest as reviews written by female students exhibiting more authentic language.

H5: Course evaluations written by male students will express more clout. Based on prior literature pertaining to a confidence gap in CS by gender, we hypothesize this trend should manifest with less expressions of clout and authority in course evaluations by female authors.

3.2.2 Hand-crafted rules

H6: Female students will write more on course evaluations and mention the instructor more often.

We hypothesize that care work will manifest in other ways beyond psycho-social variables. Specifically: female submitters will put more effort into reviews by writing more; and they will take a more individualized approach by mentioning the instructor explicitly.

[2] in our analyses, there are over twenty grade types, including + and - variants as well as credit and nocredit courses. We report A,B,C,D, and not passing grades for simplicity

We have crafted two simple measures to facilitate investigation of this hypothesis: the length of each response in number of words, and a capture of each instance of an instructor name.

3.2.3 *Topic models*

H7: There will by systematic variation in topics depending on the author's gender.

Our final analysis is exploratory using structural topic models to identify whether qualitative components of the corpora systematically vary with gender of submitter The goals of this analysis are to develop efficient means of sorting and categorizing qualitative components of course reviews.

4. DATA

Data comprise information describing enrollments in courses offered through the Computer Science (CS) Department of a private research university during the 2015-16 and 2016-17 academic years, and the entire population of formal reviews submitted by students enrolled in those courses. Reviews were administered near the end of the academic term but before the beginning of the term's official final exam period. As an incentive for submitting reviews, students were given the ability to see their final course grades a bit earlier than non-submitters.

In total these data yield 11,255 student responses from 251 courses. Courses range in character from very large introductory lecture-and-lab formats to small advanced seminars. Institutional data made available to us for analysis include each student's grade, gender, GPA, declared major (if known), and academic year. We combine these data with the corpus of reviews submitted for CS courses during the study period specified above. Approximately one-third of submitted reviews from female students, and approximately half are from undergraduates. We cannot track or identify students enrolling in multiple CS courses during the study period, however we can compute and generate response rates by grade and gender.

We limit our analysis to responses in which submitters offered a response to the review's only open-ended question. That question reads:

"What would you like to say about this course to a student who is considering taking it in the future?"

The prompt is very well aligned with our care work hypotheses, in that it specifically asks submitters to give advice to a hypothetical future student. Individual responses vary substantially in length: from a single character to over 5,964 characters (the latter equivalent to 1004 words). The mean response length is 132 characters – approximately the length of a tweet. The entire corpus of responses to this question is 300,000 words.

Additionally we analyze responses to two review prompts with five-point Likert responses: [3]

- How much did you learn from this course?
- How well did you achieve the learning goals of this course?

[3]we will limit this analysis to complete cases due to the fact that one item was not consistently administered across courses. We ignore questions pertaining to quality of instruction and focus on student learning goals.

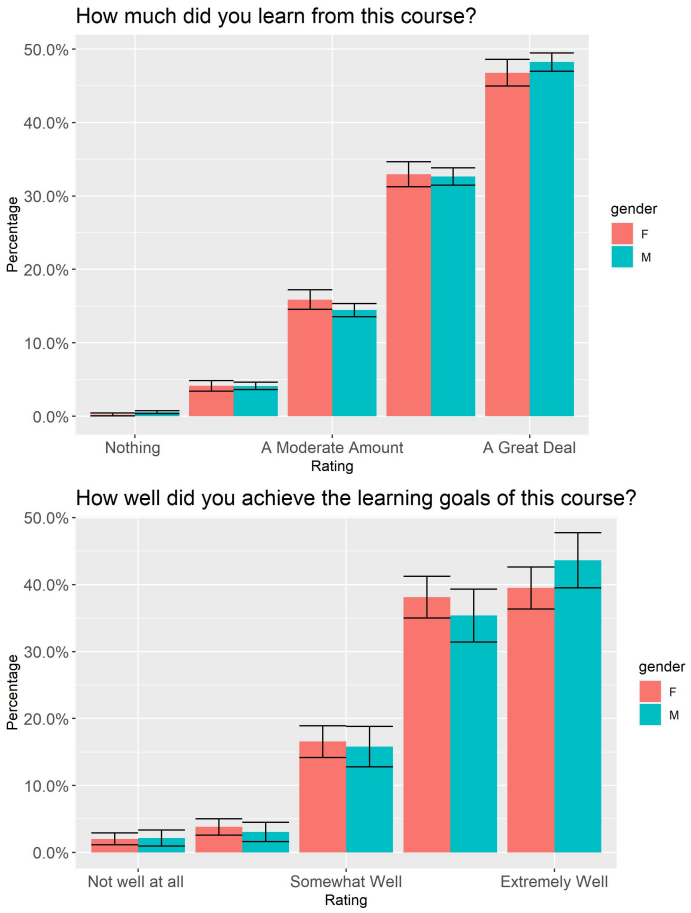

Figure 1: Responses to Likert item questions

Aggregated responses to these two prompts appear in Figure 1.

5. ANALYSES

5.1 H1: Response Rates By Gender

Rates of review submission by student gender and earned grade are reported in Figure 2. Two features are notable. First, females are more likely to submit overall. On average, females submit to 78.0% of opportunities to do so; males, 74.5% ($p<.001$).

Second, those receiving higher grades in a course are more likely to submit reviews Females receiving a grade of "A" are 3.7% more likely to respond to submit than their male counterparts. The gender submission gap is greatest among students receiving a grade of "B," with female "B" recipients 6.5% ($p<.001$)more likely to submit than males. We do not observe statistically significant gender differences in submission rates for those receiving grades below "B," however such grades represent fewer than 5% of grades in the research sample.

5.2 H2: Variation in Submission by Grade

We also examined whether review submission varied systematically grades. Figure 2 indicates a strong positive correlation between grade and likelihood of submission. A student with a grade of A or higher has an 80% chance of responding to the evaluation, while students who do not pass the course or receive credit without a grade responded approximately 50 percent of the time. Effectively,

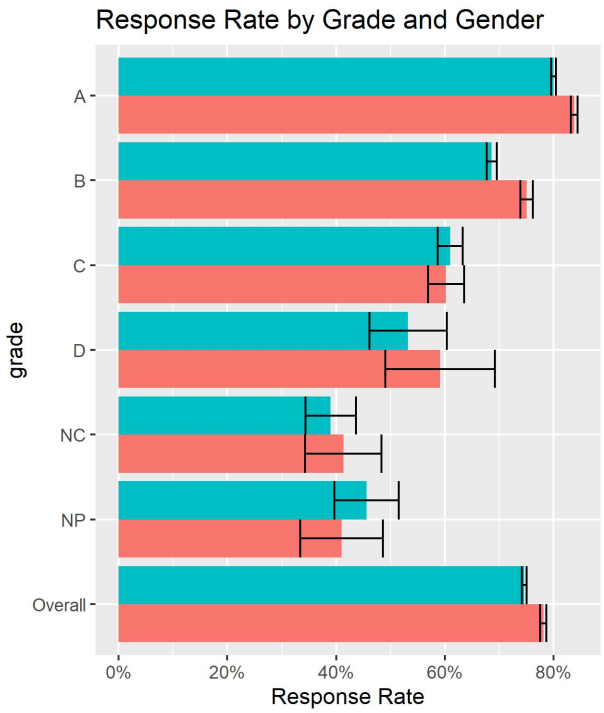

Figure 2: Survey response rates by grade

	Learning	Achievement
Male	0.01	0.08***
	(0.01)	(0.01)
Num. obs.	9002	9002
R^2 (full model)	0.15	0.18
R^2 (proj model)	0.00	0.01
Adj. R^2 (full model)	0.13	0.16
Adj. R^2 (proj model)	−0.03	−0.02
Num. groups: grade	20	20
Num. groups: evalunitid	203	203

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Gender Differences in Likert Items

this means that students who fail courses are represented by review submissions about half as often as students who excel in a courses. Differences are statistically significant with a χ^2 *statistic* of 395.37 and a p-value of less than 0.001.

5.3 H3: Reports of Learning and Goal-Meeting

We observe the proportion of students reporting having achieved the learning goals of a course *Extremely Well* by grade in Figure 3. Not surprisingly, we find a strong direct correlation with course grade, such that reported goal achievement declines with grade. What is striking is that at every grade level, there is a clear gap in reported goal achievement, with males more likely to report achievement than females earning the same grade.

We extend this analysis to see if this same pattern occurs with the question of how much students learn. Using the same specification as described in equation 1 in table 1. We see that after controlling for grades and course, males and females exhibit no differences

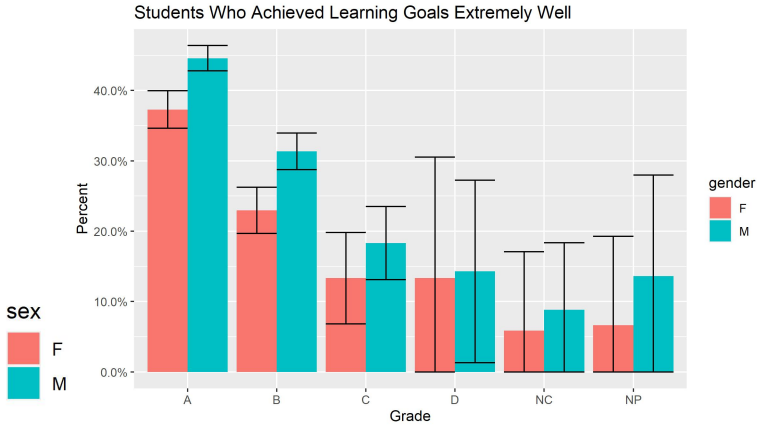

Figure 3: Percent of Students Saying they Achieved learning goals extremely well

	Tone	Analytic	Clout	Authentic
Male	−0.08**	−0.01	0.01	−0.10***
	(0.03)	(0.02)	(0.02)	(0.02)
Num. obs.	11255	11255	11255	11255
R^2 (full model)	0.07	0.05	0.05	0.04
R^2 (proj model)	0.00	0.00	0.00	0.00
Adj. R^2 (full model)	0.05	0.02	0.02	0.02
Adj. R^2 (proj model)	−0.02	−0.02	−0.02	−0.02
Num. groups: grade	20	20	20	20
Num. groups: evalunitid	251	251	251	251

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: LIWC Regressions

in self-reported measures of 'how much they learned' in a class. However, when we look at a similar question about 'achievement of learning goals', we see a stark difference. Male students are 8% points more likely than females to state they mastered the learning goals of a course. This finding suggests two concerns. First, given the similarity of these questions, we see that subtle differences in phrasing yield substantial differences in student responses. Second, female students report lower-level of mastery even after controlling for grades. Notably, these surveys are collected before students know their final grades. These perceptions may change after this information is revealed to them.

5.4 Open Text Responses

5.4.1 Psycho-social variables

graphicx

We report our analyses for ***H4*** and ***H5*** in table 2. We find modest variation by gender in how submitters describe their experience in the same course, conditional on grades. On average, submissions from males evince slightly more negative and slightly less authentic language. While these gender differences are highly significant, their magnitude is modest: on the order of a tenth of a standard deviation. Nevertheless, they are consistent with our care work hypotheses. To wit, men are somewhat more critical and less honest in their reviews than women, suggesting greater empathy and investment among female submitters.

With respect to our hypothesis around clout, we find little evidence that qualitative open-responses exhibit any significant differences in

	ProfessorName	Word Count
Male	−0.02*	−0.15***
	(0.01)	(0.03)
Num. obs.	11255	11255
R^2 (full model)	0.08	0.08
R^2 (proj model)	0.00	0.00
Adj. R^2 (full model)	0.06	0.06
Adj. R^2 (proj model)	−0.02	−0.02
Num. groups: grade	20	20
Num. groups: evalunitid	251	251

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: Handcrafted Features

submissions from males and females.

5.4.2 Handcrafted features

We report the standardized results of our analysis in table 3. We observe marginally significant differences in the frequency with which submissions from males and females mention instructor name, with women approximately two percent more likely to mention. Submissions from women are also lengthier – about .15 of a standard deviation. While modest in magnitude, these statistically significant findings comport with our care work hypotheses that female submitters approach the task of submitting reviews with more attention to specificity and investment.

5.5 Topic Models

We ran structural topic models while allowing submitter gender to vary with topic prevalence. We tuned the optimal number of topics from 2 to 50 using an exclusivity measure called FREX [2] [20]. FREX (See Equation 2) is a harmonic weighting of the frequency (F) with which a word occurs in a topic; and exclusivity (E), how frequently the word occurs in a given topic relative to others. The parameter ω corresponds to a tuning parameter of the relative importance of these features. We used the default parameter of $\omega = .7$ to favor topics that had more exclusivity.

$$FREX = (\frac{\omega}{E} + \frac{1-\omega}{F})^{-1} \quad (2)$$

Using this criterion, We found the locally optimal parameter to be 32 distinct topics. We then hand labelled each topic, observing the ten (10) statements that had the highest probability labels of that topic see Figure 4 (Top). The most common topics were comments about a course being a tutorial, a suggestion to take the course, or a positive review. The least common topics were highly specific suggestions and issues pertaining to course prerequisites. While we did not see substantial gender variation in topics overall, there are exceptions of note. First, submissions from males are more likely to talk about math prerequisites and claims that instruction was poorly organized or of poor quality. They were also more likely to discuss course organization and instruction. Submissions from females were more likely to bear topics pertaining to workload, study practices, and attendance. These patterns provide at least modest evidence that women are offering relatively more specific advice that may be relevant to larger numbers of future students.

We note an important caveat to this analysis, however. In contrast with the above studies results of the topic models presented in this section do not control for grades or course selections, thus reported gender differences in topic prevalence may be an artifact of these other factors. We attempted to model the data with all of these parameters but found the models to be degenerate.

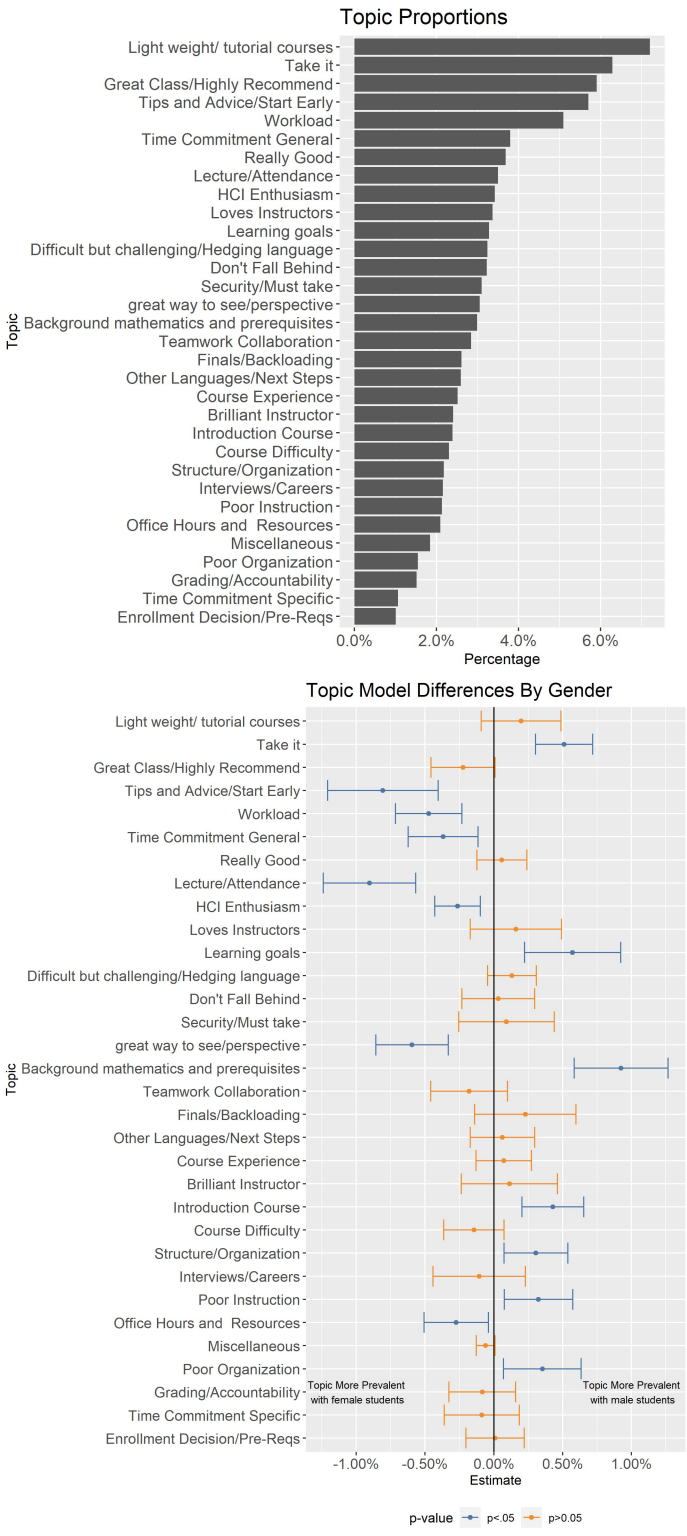

Figure 4: (Top): Topic Proportions (Bottom):Differences in Topic Prevalence by Gender

6. DISCUSSION

Even while they are controversial for evaluating instructors and instruction, course reviews are ubiquitous features of the US higher education landscape and potentially powerful tools for education data science. In the work presented here we have sought to demonstrate the promise of course reviews as a window into students' perceptions of their academic experiences and their orientation to the task of submitting evaluations. Taking advantage of archival data that included 11,255 submitted to 251 computer science courses at a single university between 2015-2017 that was linked to administrative information describing submitters' gender (M/F) and grades, we found patterned variation in who submits course reviews, and how.

In three observational studies we found that (a) women and those earning high grades were disproportionately likely to submit reviews (b) the phrasing of close-ended review prompts influenced patterns of response by gender (c) responses to qualitative review prompts differed subtly but significantly by gender, with women writing somewhat more positive, individualized, and lengthier reviews. These empirical findings comport with theoretical insights from educational social psychology and feminist social science, which suggest gender variation in how men and women perceive their own academic accomplishments and their obligations for the well-being of others.

While the empirical findings presented here are modest, they suggest the promise of leveraging course reviews for cumulative science in at least two ways.

First, we note that the inquiries presented here are based entirely on the premise that course reviews and submitter demographic information are "found" data. To the extent that virtually every US college and university possesses data such as these, we can only imagine the number and variety of insights that might be gained from parallel investigations at other schools. To a nascent field whose promise lies substantially in observing phenomena at scale, course reviews provide exceptionally promising sources of data for education data science.

Second, there is every reason to imagine that education data scientists might collaborate with school administrators to more explicitly and conscientiously instrument reviews for systematic experimental and quasi-experimental research. The basic conditions for such inquiries are already in place and sustained by established administrative rhythms: schools have offices conducting the reviews, students anticipate receiving them, and they take place multiple times a year. It is possible to imagine substantial scientific insight through the linkage review subsmissions with with administrative data describing characteristics of submitters. The initial efforts presented here provide an inkling of this promise.

As with any novel research strategy, pursuing education data science through course reviews comes with important ethical considerations regarding participant consent and responsible use. We are grateful that such discussions are already well underway nationwide [6] and we hope that our own illustrative work here might helpfully contribute to them. Indeed, addressing questions of responsible use of student data in the context of course reviews may have the additional benefit of improving the collective value of an institutional practice currently regarded with ambivalence and suspicion but that, in whatever form, will likely be part of the academic landscape for a long time.

7. REFERENCES

[1] Meredith J. D. Adams and Paul D. Umbach. 2012. Nonresponse and Online Student Evaluations of Teaching: Understanding the Influence of Salience, Fatigue, and Academic Environments. *Research in Higher Education* 53, 5 (8 2012), 576–591. DOI: http://dx.doi.org/10.1007/s11162-011-9240-5

[2] Edoardo M Airoldi and Jonathan M Bischof. 2012. A Poisson convolution model for characterizing topical content with word frequency and exclusivity. *arxiv.org* (2012). https://arxiv.org/abs/1206.4631http://arxiv.org/abs/1206.4631

[3] Christine Alvarado, Yingjun Cao, and Mia Minnes. 2017. Gender Differences in Students' Behaviors in CS Classes throughout the CS Major. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. ACM, New York, NY, USA, 27–32. DOI: http://dx.doi.org/10.1145/3017680.3017771

[4] A.J. Alvero, Noah Arthurs, anthony lising antonio, Benjamin W. Domingue, Ben Gebre-Medhin, Sonia Giebel, and Mitchell L. Stevens. 2020. AI and Holistic Review. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 200–206. DOI: http://dx.doi.org/10.1145/3375627.3375871

[5] David N. Beede, Tiffany A. Julian, David Langdon, George McKittrick, Beethika Khan, and Mark E. Doms. 2011. Women in STEM: A Gender Gap to Innovation. *SSRN Electronic Journal* (8 2011). DOI: http://dx.doi.org/10.2139/ssrn.1964782

[6] Michael Brown and Carrie Klein. 2020. Whose Data? Which Rights? Whose Power? A Policy Discourse Analysis of Student Privacy Policy Documents. *The Journal of Higher Education* (2020), 1–30.

[7] Shelley J. Correll. 2004. Constraints into Preferences: Gender, Status, and Emerging Career Aspirations. *American Sociological Review* 69, 1 (2 2004), 93–113. DOI: http://dx.doi.org/10.1177/000312240406900106

[8] Erica DeFrain and Erica DeFrain. 2016. An Analysis of Differences in Non-Instructional Factors Affecting Teacher-Course Evaluations over Time and Across Disciplines. (2016). https://repository.arizona.edu/handle/10150/621018

[9] Paula England. 2005. Emerging Theories of Care Work. *Annual Review of Sociology* 31, 1 (8 2005), 381–399. DOI: http://dx.doi.org/10.1146/annurev.soc.31.041304.122317

[10] Nancy Folbre. 1995. "Holding hands at midnight": The paradox of caring labor. *Feminist Economics* 1, 1 (3 1995), 73–92. DOI:http://dx.doi.org/10.1080/714042215

[11] Maarten Goos and Anna Salomons. 2017. Measuring teaching quality in higher education: assessing selection bias in course evaluations. *Research in Higher Education* 58, 4 (6 2017), 341–364. DOI: http://dx.doi.org/10.1007/s11162-016-9429-8

[12] Nan Hu, Jie Zhang, and Paul A. Pavlou. 2009. Overcoming the J-shaped distribution of product reviews. (10 2009). DOI: http://dx.doi.org/10.1145/1562764.1562800

[13] Neneh Kowai-Bell, Rosanna E. Guadagno, Tannah Little, Najean Preiss, and Rachel Hensley. 2011. Rate My Expectations: How online evaluations of professors impact students' perceived control. *Computers in Human Behavior* 27, 5 (9 2011), 1862–1867. DOI:

http://dx.doi.org/10.1016/J.CHB.2011.04.009

[14] Cong Li and Xiuli Wang. 2013. The power of eWOM: A re-examination of online student evaluations of their professors. *Computers in Human Behavior* 29, 4 (7 2013), 1350–1357. DOI: http://dx.doi.org/10.1016/J.CHB.2013.01.007

[15] Jin Liu and Julie Cohen. 2020. Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods | EdWorkingPapers. (2020). https://www.edworkingpapers.com/ai20-239

[16] Anna May, Johannes Wachs, and Anikó Hannák. 2019. Gender differences in participation and reward on Stack Overflow. *Empirical Software Engineering* 24, 4 (8 2019), 1997–2019. DOI: http://dx.doi.org/10.1007/s10664-019-09685-x

[17] Arunachalam Narayanan, William J. Sawaya, and Michael D. Johnson. 2014. Analysis of Differences in Nonteaching Factors Influencing Student Evaluation of Teaching between Engineering and Business Classrooms. *Decision Sciences Journal of Innovative Education* 12, 3 (7 2014), 233–265. DOI:http://dx.doi.org/10.1111/dsji.12035

[18] JW Pennebaker, RL Boyd, K Jordan, and K Blackburn. 2015. The development and psychometric properties of LIWC2015. (2015). https://repositories.lib.utexas.edu/handle/2152/31333

[19] Katie Redmond, Sarah Evans, and Mehran Sahami. 2013. A large-scale quantitative study of women in computer science at Stanford University. In *Proceeding of the 44th ACM technical symposium on Computer science education - SIGCSE '13*. ACM Press, New York, New York, USA, 439. DOI:http://dx.doi.org/10.1145/2445196.2445326

[20] J Reich, DH Tingley, J Leder-Luis, and M Roberts. 2014. Computer-assisted reading and discovery for student generated text in massive open online courses. (2014). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2499725

[21] Lauren A Rivera and András Tilcsik. 2019. Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review* 84, 2 (2019), 248–274.

[22] Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science* 58, 4 (10 2014), 1064–1082. DOI:http://dx.doi.org/10.1111/ajps.12103

[23] Linda J. Sax, Shannon K. Gilmartin, and Alyssa N. Bryant. 2003. Assessing response rates and nonresponse bias in web and paper surveys. (2003). DOI: http://dx.doi.org/10.1023/A:1024232915870

[24] T Sliusarenko, LH Clemmensen International . . . , and Undefined 2013. 2013. Text Mining in Students' Course Evaluations. *pdfs.semanticscholar.org* (2013). https://pdfs.semanticscholar.org/cb02/b880ef86371461b3ebe46d2f8c293b43c7a2.pdf

[25] Philip Stark, Kellie Ottoboni, and Anne Boring. 2016. Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness. *ScienceOpen Research* (2016). DOI:http://dx.doi.org/10.14293/s2199-1006.1.sor-edu.aetbzc.v1

[26] Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson Murphy-Hill, Chris Parnin, and Jon Stallings. 2017. Gender differences and bias in open source: pull request acceptance of women versus men. *PeerJ Computer Science* 3 (5 2017), e111. DOI: http://dx.doi.org/10.7717/peerj-cs.111

[27] Bob Uttl, Carmela A. White, and Daniela Wong Gonzalez. 2017. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation* 54 (9 2017), 22–42. DOI: http://dx.doi.org/10.1016/J.STUEDUC.2016.08.007