4

USING STRUCTURAL TOPIC MODELLING TO ESTIMATE GENDER BIAS IN STUDENT EVALUATIONS OF TEACHING

Marshall A. Taylor, Ya Su, Kevin Barry, and Sarah A. Mustillo

Introduction

Student evaluations of teaching (SETs) are widely used in higher education (HE) institutions as an important measure of instructors' teaching qualities. Many institutions of HE in the US and throughout the world rely on SETs for both formative and summative purposes (Zabaleta, 2007; Spooren, Brockx, and Mortelmans, 2013). SETs provide useful information that helps instructors improve their courses and assist administrators in evaluating instructors for promotions and merit raises. However, many recent studies show that student perceptions of instructor effectiveness may be based in part on teacher social characteristics rather than solely on course and teaching quality (Basow and Montgomery, 2005; Sprague and Massoni, 2005; Basow, Phelan, and Capotosto, 2006; Boring, Ottoboni, and Stark, 2016). Such research demonstrates that seemingly objective evaluation systems are not simply neutral instruments to measure merits; these systems could be potential sources of (many times unintended) bias that may produce and exacerbate social inequality in HE.

The growing body of research on detecting biases in SETs focuses primarily on disparities in numerical ratings (MacNell, Driscoll, and Hunt, 2015; Wagner, Rieger, and Voorvelt, 2016; Boring, 2017; Rivera and Tilcsik, 2019). These studies provide valuable insights into revealing how systematic biases in SETs negatively impact certain social groups. However, it is likely that there are other sources of bias contributing to the unevenly distributed opportunities, resources and awards in SETs, but are hard to capture in numeric scores. In this regard, some studies found that students have gender-specific or race-specific expectations in the criteria they use to evaluate their instructors in the context of open-ended qualitative questions (Sprague and Massoni, 2005; Basow, Phelan, and Capotosto, 2006).

DOI: 10.4324/9781003138785-5

Recent advances in computational text-analytic tools have made it possible for researchers to expand on these qualitative inquiries in the era of 'big data'. These methods provide the opportunity to map 'cultural environments' (Bail, 2014). Drawing on more than 172,000 open-ended student comments from 2013 through 2015 at a private research university as a case illustration, we demonstrate how to use structural topic modelling (STM) to uncover one particular type of bias – gender bias – in SET comments at previously unimaginable scales. We also discuss the limitations and promises of our approach. We believe that findings from our study will help administrators develop more effective evaluating tools as well as help decision-makers attune to potential biases in evaluating instructors in HE and beyond.

Purpose and context of the analysis

This project originates from a committee on the evaluation of teaching at a North American research university. The committee reviewed the holistic evaluation of teaching that had been adopted seven years earlier at the university. The components of that evaluation process were: course design, implementation, evaluation of student work, and student perceptions of their learning experience.

In the two-year process of reviewing the components of this system and making recommendations for change, the committee assessed the existing student feedback instrument and made recommendations for changes to the instrument and the reporting of results. As part of that process the committee investigated the utilisation of open-ended student comments to determine if these comments themselves would result in new sources of bias in the evaluation process.

This study was one part of a broad review of the holistic evaluation of teaching that had been occurring at the university for nearly ten years. In the existing system for student feedback, students' open-ended comments were only available to the instructor and were not included as part of the summative evaluation of teaching. To gauge sentiment on the use of comments in the summative evaluation, both students and faculty were surveyed on the topic of adding access to comments in that process. A large majority of both undergraduate and graduate students were in favour of contributing feedback for the evaluation process and for their comments to be included. For faculty, there was a slight majority in favour of the use of student comments. The results of that broad review were communicated to the entire faculty in an extensive report that included recommendations to the provost. Those recommendations included adding access to student comments to support formative development of instructors in their departments and access to those comments at all levels of the summative evaluation process. That recommendation was, in part, based on the information that was provided by this analysis of student comments for gender bias. There were also recommendations for modifications to the student feedback instrument, for reporting of quantitative student feedback, and for a more uniform implementation of peer observation/evaluation as part of the summative evaluation process.

Regarding the use of student comments, the recommendations came with several stipulations, including:

- the student comments were not to be a new category of evaluation but would be used as a source of additional information in the student feedback category;
- the use of direct quotes was to be prohibited in teaching statements of the candidate for renewal, promotion or tenure, and in departmental reports and chairs' letters;
- candidates should be encouraged to reflect on both quantitative and textual feedback in their teaching statements;
- comments should be accessible to all involved in the evaluation process rather than having the task of summarising being limited to a single evaluator; and
- that decision-makers would be coached on how to identify themes in the comments.

All recommendations were adopted by the provost.

This chapter is devoted to the discussion of that investigation and the techniques used to examine the student comments to uncover evidence of bias – specifically, gender bias. Though we investigate gender biases, this chapter is meant to function as a case illustration for how to use STM to analyse SETs – not necessarily as a contribution to the literature on gender biases in SETs. Readers interested in the current state of this literature are encouraged to see work cited here (Centra and Gaubatz, 2000; Arbuckle and Williams, 2003; MacNell, Driscoll, and Hunt, 2015; Boring, 2017; Rivera and Tilcsik, 2019).

Assessing gender bias in student evaluations

Student evaluation data

The investigations are based on a database of de-identified SETs collected across five consecutive semesters from Fall 2013 to Fall 2015. The dataset was limited to undergraduate courses (100- to 400-level courses) taught by members of the regular faculty (e.g., this did not include data on graduate student instructors, adjunct professors, etc.) with enrollments of at least five students. Data were collected via an online system that allowed students to complete their evaluations at a time and place of their own choosing during the final days of each semester, excluding final exams. Participation was voluntary, although students were incentivised to complete SETs by offering early access to final grades. Analysis of these data for research purposes was approved by the institutional review board at the university.

In the section of open-ended responses, students were given three prompts guiding their written feedback. Students were asked to *identify what they perceive to be the greatest strengths of this instructor's teaching*, to *identify areas where this instructor could improve his/her teaching*, and to *comment on how well the activities, readings, lectures, and assignments helped them to learn in this course*. In this study, the authors focused on students' responses

concerning instructor strengths and weaknesses derived from the first two questions. In total, the data included 93,740 comments on instructor strengths and 78,434 comments on instructor weaknesses after text preprocessing and observation deletion due to data sparsity or missingness across the independent and control variables.

Measures

The internal analysis looked at a range of course-, instructor-, and student-level sources of bias in the open-ended responses. The focus of this case illustration is on student and instructor gender differences. However, we share results of the gender differences after controlling for these other factors. Specifically, in addition to the primary independent variables (dummy variables for instructor and student gender – where 0 = identifies as male and 1 = identifies as female – and an interaction term between the two), we also control for the following instructor-level factors: race/ethnicity/nativity, age, and academic rank. The student-level control variables were race/ethnicity/nativity, age, and student level (e.g., senior, junior, etc.). The course-level factors were level (i.e., 100-level, 200-level, etc.), enrollment, students' reasons for taking the course, students' perceived course difficulty, time spent on the course outside of class, expected grade, academic division (i.e., Social Sciences, Humanities, etc.), and academic semester (i.e., Fall 2013, Spring 2014, etc.).

The dependent variables are derived from the STM procedure, which are detailed in the following section.

Carrying out the analysis

Structural topic modelling

Analysing comments across more than 172,000 open-ended survey responses demands a method for reducing the complexity of the comments to their 'core themes'. To this end, students' qualitative comments were analysed using STM (Roberts et al., 2013). STM can be thought of as a method for uncovering the underlying themes (or 'topics') across a set of texts. The goal is to extract the latent topic structure of the overall text dataset (the corpus) by finding the topics (or content themes) that best account for the co-occurrence of words in a document. We can then treat the predicted proportion of words in each document that was pulled from each topic (i.e., document-topic probabilities) as dependent variables whose values can be observed across levels of various independent/control variables. We treated comments regarding instructor strengths and weaknesses as separate text datasets (also known as 'corpora') as the goal was to create separate topic solutions for each set of comments.

Getting the text data ready

In traditional text pre-processing fashion, all capitalisation, punctuation,[1] numbers, excess whitespace, standard English 'stop words' (e.g., non-discriminant words such

the articles 'the', 'an' and 'a'),[2] custom stop words,[3] and words with fewer than three characters were removed. Words were also stemmed using the Porter2 English stemming algorithm (Porter, Boulton, and Macfarlane, 2002). For example, words like lecture, lectures, lecturing, lectured, and so on were reduced to 'lectur' so that a single stem could capture multiple variations of the same word. Additionally, any words that appeared in only 1% or fewer responses were removed from the analysis so as to sort out overly idiosyncratic terms. Any comments with missing data regarding our independent/control variables were removed prior to generating the topics. Net of all text pre-processing and listwise deletion, the instructor strength corpus used to construct strength topics consisted of 210 unique words across 14,390 total words, 93,740 comments, 11,135 students, and 1,067 instructors. The instructor weakness corpus used to create weakness topics contained 226 unique words across 17,360 total words, 78,434 comments, 10,732 students, and 1,066 instructors.

Predicting topic engagement

Once topics were computed and interpreted, we examined students' qualitative evaluations of instructors' 'strength topics' (hereafter just 'strengths') and 'weakness topics' (hereafter just 'weaknesses') using a series of multi-level linear regression models. The strength and weakness topics were computed at the level of the individual, course-specific evaluation. As such, we estimated three-level ordinary least squares regression models with evaluations nested in course and course nested in instructor.[4]

Project findings

Analyses that use topics as dependent variables in regression models are twofold: first, estimate and interpret the topics; then, perform the regressions. As such, in what follows, we first describe how we used STM to estimate and interpret the strengths and weaknesses articulated in the student responses. We then present the results of the regression models to see how engagement with particular strengths and weaknesses varies depending on instructor and student gender net of other characteristics.

Topic estimation and interpretation

We computed topics separately for student responses regarding instructor strengths and weaknesses and ran two independent STM analyses. A series of models with different numbers of topics (ranging from 5 to 50, in increments of 5) were computed for each of the two corpora, conditioning the topic model solutions on the independent and control variables. Some model fit statistics favoured a 5-topic solution for the two corpora; other statistics favoured 15- or 20-topic solutions. We chose to focus our content validation efforts on a 10-topic solution for each corpus, striking a balance between the statistical recommendations. To verify that

the strength and weakness topics made substantive sense, we compared the 10-topic solutions to the 5-topic solutions and concluded that the 10-topic solutions struck the most consistent balance between identifying the uniqueness of each strength/weakness while retaining an appropriate level of generality. It is important to keep in mind, however, that topic model results can be sensitive to a range of analytical factors; as such, the topics reported here should be seen as one possible thematic representation of these open-ended responses.

The 10 strength and weakness topics identified through the topic model analyses are listed in Tables 4.1 and 4.2, respectively, along with the top 10 most probable and 10 most distinguishing terms for each strength/weakness. The most probable words are those that would have the highest probability of being drawn at random from all the words associated with that strength/weakness. The words that were most distinguishing were highly associated with a particular strength/weakness and poorly associated with other strengths/weaknesses.

All strength and weakness variables are probabilities, and are therefore constrained between the unit interval [0,1]. Figure 4.1 plots the overall expected proportions of the strength and weakness corpora that are allocated to each strength topic (left panel) and weakness topic (right panel), respectively.

Examining discursive gender biases

Having identified strength topics and weakness topics used by students to describe their instructors, we can now address our main question of interest: *Does the distribution of strength and weakness topics vary across gender categories?* Given that we estimate 20 total regression models, we focus here on just the instructor and student gender effects via a series of predicted probabilities. All other variables are controlled for in all models.[5]

Figure 4.2 presents a series of predicted probabilities for each strength by instructor gender and student gender. Graphs that are not opaque represent a statistically significant interaction between instructor gender and student gender ($p < .05$). As the figure shows, the extent to which a male versus female educator (instructor) is appraised as having a particular strength varies depending on the student's gender for 8 of the 10 strengths. Among the strengths with a statistically significant interaction between instructor and student gender, both male and female instructors are most likely to be praised for making the material easy to understand, regardless of whether a male or female student wrote the comment. A SET for a class taught by a female instructor, for example, has a 0.12 probability of addressing how she makes the material easy to understand regardless of whether it is a female or male student doing the writing, while a SET for a male-taught class has a 0.11 or 0.12 probability of being about this strength depending on whether it is a male or female student evaluation. The difference in these predicted probabilities is substantively quite small, with the largest difference in probabilities coming from the 'Facilitates class discussion' and 'Availability' strengths, where a SET for a female-taught class and written by a female student has 0.024 and 0.027 higher probabilities of being

TABLE 4.1 Predominant strengths articulated by students

Makes material easy to understand	*Knowledge of and passion for source material*	*Nice, caring, and excited about student learning*	*Availability*	*Enthusiasm and humor*	*"One of the Best Teachers I've Ever Had"*	*Excels in communication and maintains engagement in lectures*	*Facilitates class discussion*	*Uses real-world examples*	*Encourages critical and creative thinking*
Terms with the highest marginal probability within each strength									
understand	knowledg	student	help	interest	class	great	class	exampl	student
lectur	subject	teach	alway	make	one	good	discuss	materi	think
materi	passion	know	question	class	teacher	job	like	use	work
organ	materi	realli	will	engag	cours	engag	also	keep	way
explain	matter	care	class	materi	best	topic	realli	class	class
concept	person	well	outsid	fun	learn	class	read	engag	strength
well	extrem	can	answer	enthusiasm	take	thing	enjoy	relat	learn
clear	experi	want	time	made	everi	prof	appreci	real	cours
prepar	talk	clear	student	humor	much	excel	thought	lot	greatest
easi	hes	communic	avail	abl	semest	lectur	just	world	differ
Terms with the highest marginal lift within each strength*									
follow	matter	tell	offic	sens	—	job	opinion	real	greatest
note	hes	clariti	hour	humor	—	excel	allow	world	strength
lab	subject	can	answer	entertain	ive	thing	read	applic	critic
powerpoint	intellig	amaz	avail	knew	ever	great	first	life	encourag
slide	knowledg	care	question	enthusiasm	favorit	good	discuss	relat	particip
easi	obvious	communic	ask	interest	best	hard	day	bring	develop

(*Continued*)

TABLE 4.1 (Continued)

Makes material easy to understand	*Knowledge of and passion for source material*	*Nice, caring, and excited about student learning*	*Availability*	*Enthusiasm and humor*	*"One of the Best Teachers I've Ever Had"*	*Excels in communication and maintains engagement in lectures*	*Facilitates class discussion*	*Uses real-world examples*	*Encourages critical and creative thinking*
practic	field	teach	meet	fun	one	prof	listen	use	promot
problem	passion	everyth	will	excit	teacher	topic	peopl	attent	think
exam	guy	expect	outsid	funni	awesom	kept	come	relev	feedback
detail	experi	know	fair	connect	semest	especi	thought	effort	activ
Semantic coherence for each strength									
−105.994	−131.991	−109.314	−98.551	−122.526	−110.905	−118.959	−111.285	−117.659	−104.777

Note: Missing cell entries (indicated with "---") identified the university used in the analysis and were therefore removed from the table. The terms with the highest marginal probability can be interpreted as those terms most likely to be associated with that strength; the terms with the highest marginal lift can be interpreted as those terms most distinctive to the strength relative to the other strengths.

* For the formula behind the calculation of lift, see Taddy, 2013, p. 757 and Sievert and Shirley, 2014, p. 66.

TABLE 4.2 Predominant weaknesses articulated by students

Hard to follow	*Not engaging*	*Needs more opportunities for practice*	*Too fast and dry*	*"Can't Think of a Weakness"*	*Too Much work and poor feedback on that work*	*Poor time management*	*"I Think" and "I Feel That" the course was*	*Poorly implemented and disorganized readings*	*Unhelpful and unapproachable*
Terms with the highest marginal probability within each weakness									
lectur	class	materi	sometim	teach	assign	time	learn	read	student
hard	discuss	problem	littl	think	grade	class	think	like	question
note	none	test	can	improv	work	talk	class	felt	know
slide	make	exam	bit	need	semest	less	just	realli	answer
inform	engag	homework	get	cours	cours	long	better	didnt	ask
difficult	mayb	exampl	topic	good	project	wish	noth	much	even
use	interest	help	thing	see	lab	end	like	alway	dont
confus	student	understand	focus	great	paper	day	thing	never	hour
organ	come	concept	seem	prof	expect	everi	book	differ	help
follow	keep	practic	certain	cant	one	spend	well	class	want
Terms with the highest marginal lift within each weakness*									
note	engag	practic	slow	area	project	minut	best	wasnt	offic
powerpoint	interest	problem	bit	improv	receiv	spent	learn	part	ask
slide	activ	review	dri	teach	week	spend	noth	read	answer
follow	none	move	littl	style	final	start	book	didnt	question
board	speak	exam	focus	job	feedback	long	anyth	quizz	hour
onlin	particip	studi	sometim	prof	due	appreci	done	sinc	look

(*Continued*)

TABLE 4.2 (Continued)

Hard to follow	*Not engaging*	*Needs more opportunities for practice*	*Too fast and dry*	*"Can't Think of a Weakness"*	*Too Much work and poor feedback on that work*	*Poor time management*	*"I Think" and "I Feel That" the course was*	*Poorly implemented and disorganized readings*	*Unhelpful and unapproachable*
inform	subject	prepar	can	cant	paper	everyon	outsid	never	even
explan	pretti	test	certain	teacher	grade	talk	sure	suggest	said
hard	bore	goe	lost	need	lab	time	show	direct	know
organ	discuss	homework	topic	general	assign	less	rather	differ	want
Semantic coherence for each weakness									
−109.96083	−144.22958	−100.77706	−118.47108	−128.45162	−111.74025	−132.06236	−114.68981	−93.17845	−98.82477

Note: The terms with the highest marginal probability can be interpreted as those terms most likely to be associated with that weakness; the terms with the highest marginal lift can be interpreted as those terms most distinctive to the weakness relative to the other weaknesses.

* For the formula behind the calculation of lift, see Taddy, 2013, p. 757 and Sievert and Shirley, 2014. p. 66.

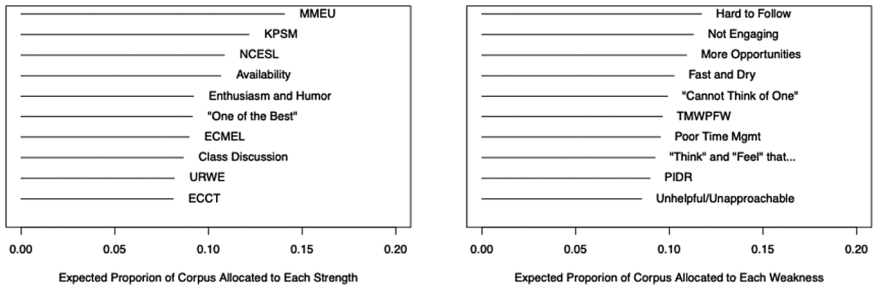

FIGURE 4.1 Expected overall proportions for each strength/weakness

Note: The expected proportions for each strength (left panel) refer to how many of the total words in the instructor strengths corpus are accounted for by that strength. The expected proportions for each weakness (right panel) are interpreted the same way, but with reference to the instructor weaknesses corpus.
MMEU = Makes material easy to understand; KPSM = Knowledge of and passion for source material; NCESL = Nice, caring, and excited about student learning; ECMEL = Excels in communication and maintains engagement in lectures; URWE = Uses real-world examples; ECCT = Encourages critical and creative thinking; TMWPFW = Too much work and poor feedback on that work; PIDR = Poorly implemented and disorganised readings.

about those respective strengths than a SET for a male-taught class and written by a male student.

We then re-ran these analyses with the 10 weakness topics as the dependent variables. Figure 4.3 presents a series of predicted probabilities for each weakness by instructor gender and student gender. Three of the perceived weaknesses – being *'Unhelpful and unapproachable'*, giving *'Too much work and poor feedback on that work'* and *'Poor time management'* – did not display a statistically significant interaction between instructor and student gender, suggesting that differences in how male and female instructors are evaluated in relation to these weaknesses do not vary much by the gender of the student providing the evaluation. The largest predicted probabilities come from the *'Hard to follow'* weakness, where a SET written by a female student for a male-taught class has about a 0.11 probability of being about this weakness (but with all other instructor–student combinations hovering around a 0.10 probability). The largest difference in probabilities comes from the perceived lack of a weakness between male and female students evaluating male instructors. SETs from male students addressed to male instructors have a 0.6 percentage point higher chance of noting no weaknesses than SETs from female students addressed to male instructors.

Conclusion

We found that the extent to which a SET addresses a particular strength or weakness varies by instructor and student gender. The extent to which a male instructor versus a female instructor is appraised with a strength varies by student gender for 8

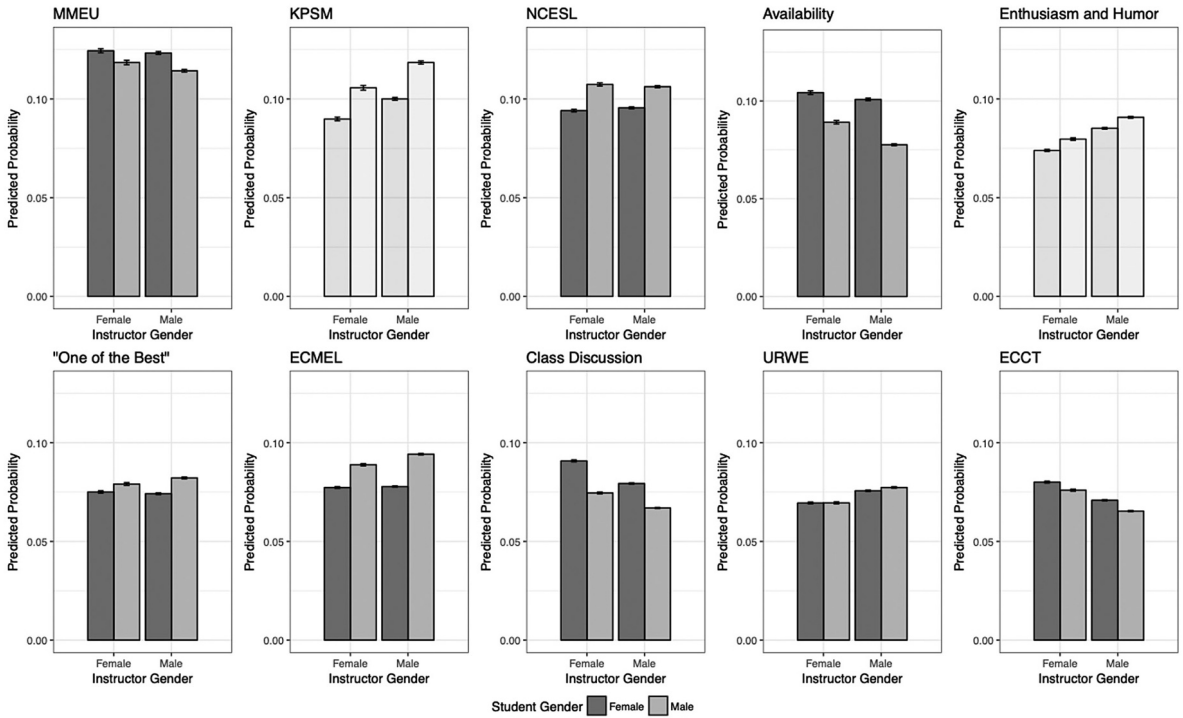

FIGURE 4.2 Predicted probabilities of SET addressing a strength in relation to Instructor gender, by student gender

Note: Vertical bars are 99% confidence intervals. Predicted probabilities and their confidence intervals are derived from taking the logistic function of the predicted log odds. Opaque graphs indicate a statistically non-significant interaction between instructor gender and student gender ($p \geq .05$). Strengths are arrayed from left to right in the order presented in Table 4.1. See the caption to Figure 4.1 for interpreting the topic acronyms.

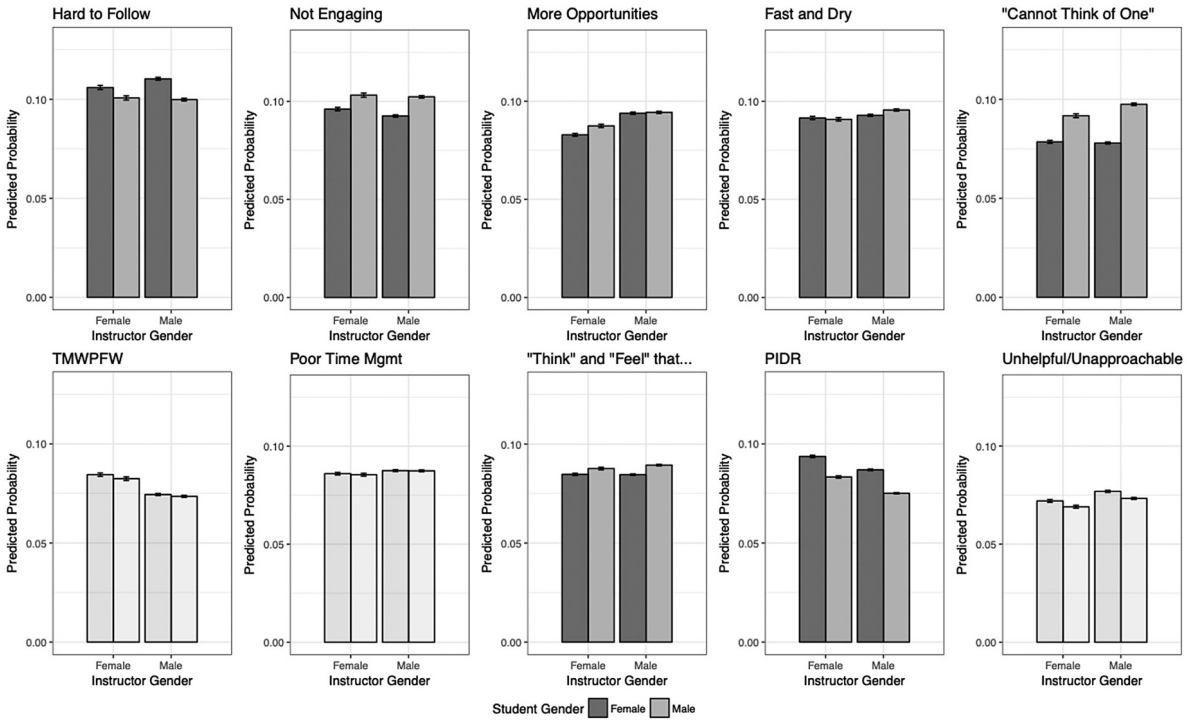

FIGURE 4.3 Predicted probabilities of SET addressing a weakness in relation to Instructor gender, by student gender

Note: Vertical bars are 99% confidence intervals. Predicted probabilities and their confidence intervals are derived from taking the logistic function of the predicted log odds. Opaque graphs indicate a statistically non-significant interaction between instructor gender and student gender ($p \geq 0.05$). Weaknesses are arrayed from left to right in the order presented in Table 4.2. See the caption to Figure 4.1 for interpreting the topic acronyms.

of 10 strengths: *Makes material easy to understand; Nice, caring, and excited about student learning*; *Availability*; *One of the best*; *Excels in communication and maintains engagement in lectures*; *Class discussion*; *Uses real-world examples;* and *Encourages critical and creative thinking*. This instructor-student gender interaction effect is also present for 7 of 10 weaknesses: *Hard to follow*; *Not engaging*; *More opportunities*; *Fast and dry*; *Cannot think of one*; *Think and 'feel' that…*; and *Poorly implemented and disorganised readings*. Our findings suggest that student and instructor gender play roles in shaping open-ended evaluations of teaching. Gender-based stereotypes may affect the type of feedback given in course evaluations. In this sense, it is fruitful for future research to utilise this approach for further investigations on or beyond the realm of gender. The (perceived) race, ethnicity and national origin of instructors, for example, can invoke certain stereotypes that shape evaluations of teaching. We encourage future research to unpack a wide range of social characteristics of instructors and students in assessing SETs.

The work described in this chapter is related to a project that was undertaken to inform the decision about adding student comments to the information available for renewal, promotion and tenure cases. In a case where the quantitative results of student feedback demonstrate low levels of bias for any single identity, there was a concern about introduction of bias by adding comments to the evaluation materials. Based on the results of the study and the desire to provide a richer source of information to inform cases that are not clear-cut, the decision to add comments was recommended and approved. At this time, there has not been a study of the frequency or impact of including the comments.

Like any method, there are some limitations with using topic modelling to analyse SETs. One limitation concerns the inherently subjective and interpretive nature of the topic modelling procedure. Like any dimension reduction technique, deciding on the appropriate number of latent topics to retain for analysis largely boils down to the solution that best passes the eye test. As such, it is an empirical question as to whether the general results reported here would hold with more fine-grained topics. The results reported here, then, are derived from one possible thematic representation of the comments.

Another limitation concerns effect sizes. Namely, how do we know a 'large' group difference in topic engagement when we see one? Is the difference between a 0.12 probability that a SET written by a male student and for a course by a female instructor addresses the '*Makes material easy to understand*' strength and a 0.11 probability that a SET written by a male student and for a course by a male instructor addresses this same strength a big difference? What about the difference between a 0.11 probability and a 0.22 probability? It might be tempting to interpret a 0.01 difference in probabilities as a very small difference given the 0–1 range – and perhaps that is the correct answer. But, to our knowledge, there is very little to no agreement on baselines for assessing what a substantively small vs. moderate vs. large effect is – at least within the context of the social sciences. Future research should address this question in more detail, and SET researchers using STM should think critically about the question of substantive significance when they interpret results.

In summary, this study provides new methodological tools for uncovering potential biases in SETs. Previous research suggested that different modes of evaluation are not simply neutral evaluative tools (Rivera and Tilcsik, 2019). Instead, they have different impacts on the extent to which biases are reflected in instructors' performance evaluations. Insofar as SETs are used to evaluate performance and distribute rewards in HE, utilising a wide array of evaluative tools to detect potential biases is fruitful to improve workplace inequality. We encourage future research to examine numeric ratings and qualitative evaluations jointly to detect multiple sources of biases, and to design equitable evaluation tools in measuring instructors' teaching effectiveness.

Practical tips

This type of computational text analysis poses some learning curves. As such, our most important advice is familiarisation with at least one of the more popular and flexible platforms for carrying out an analysis such as this: R or Python. Sometimes, though, a particular method is implemented in one of these programming languages but not the other. This means that analysts may want to specialise in one particular programming language but be comfortable enough to selectively use the other when they need a particular task done that their primary language is not equipped to handle.

Another element to consider is the size of the data and the available computing resources. This project consisted of over 172,000 SETs, each accompanied by a large set of student-, instructor-, and course-level covariates. Further, STM uses what is known as an 'expectation-maximization algorithm' (EM algorithm) to estimate the topics, which is an iterative method where the topics are estimated many times until it finds the model estimates that best maximise the likelihood of the observed data. The EM algorithm tends to take longer to converge as the number of documents increases. With a corpus as large as ours and a time-intensive procedure such as STM, we decided to use high performance computing (HPC) resources. Using these types of resources means writing up the necessary code on the analyst's local machine, checking that the code works with small samples of texts either using the computing resources on that local machine or by patching into a front-end compute node in the HPC system (if one exists), and then submitting that code and the necessary data as a 'job' into a 'queue' where the code will be run on another compute node (or set of nodes). Many research-oriented institutions of HE have HPC resources.

A benefit of using HPC resources is that a job can typically be run faster than it would be using local computing resources – depending on the parameters of the job set by the analyst. This also means that the analyst's local computing resources (e.g., their machine's RAM) are freed up for other tasks. There might be a small learning curve associated with using HPC, such as learning the basics of shell scripting, but institutions usually provide tutorials or training to orient analysts on these matters.

Lastly, we want to emphasise that topic modelling – or any other type of computational text analysis – should not be treated as a way to purely automate the analysis

of meaning. Topic modelling is often conceptualized as a form of 'unsupervised machine learning' for texts; however, these types of unsupervised learning tools should be treated as automating pattern detection, not (at least in isolation) pattern interpretation (Nelson, 2020, p. 34). While we used various statistical measures to help guide our choice of strength/weakness topics, this was never done in isolation from a careful, qualitative reading of SETs that were representative of each topic. These selective readings were important for giving the topics meaning. As Nelson (2020, p. 34) notes, computers may be able to find patterns in language, but they cannot (yet) interpret them. We echo this position here.

Notes

1 Dashes internal to a word (e.g. 'micro-level') were retained.
2 We used the SMART stoplist (Lewis et al., 2004), which contains 571 words for removal.
3 The custom stop words were the following: 'professor', 'instructor' 'faculty' and a selection of university-specific terms.
4 With over 172,000 comments across the two samples, we found we were able to ignore student-level clustering and simply nest SETs in course. Also, the dependent variables – the topic probabilities, bound between 0 and 1 – were logit-transformed prior to estimating the models. Technical details on the estimation strategy are available from Marshall A. Taylor on request.
5 All predicted probabilities are adjusted predictions at the means (Williams, 2012): i.e., continuous and categorical controls are set at their means.

References

Arbuckle, J. and Williams, B. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles*, 49(9–10), 507–516.

Bail, C. (2014). The cultural environment: measuring culture with big data. *Theory and Society*, 43(3–4), 465–482.

Basow, S. and Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education*, 18(2), 91–106.

Basow, S., Phelan, J. and Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly*, 30(1), 25–35.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145, 27–41.

Boring, A., Ottoboni, K. and Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. doi: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Centra, J. and Gaubatz, N. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education*, 71(1), 17–33.

MacNell, L., Driscoll, A. and Hunt, A. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40, 291–303.

Nelson, L. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49(1), 3–42.

Porter, M., Boulton, R. and Macfarlane, A. (2002). *The English (porter2) stemming algorithm*. Available from: https://snowballstem.org/algorithms/english/stemmer.html (Accessed 28 January 2021).

Rivera, L. and Tilcsik, A. (2019). Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review*, 84(2), 48–74.

Roberts, M., Stewart, B., Tingley, D. and Airoldi, E. (2013). *The structural topic model and applied social science*. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*, Lake Tahoe, Utah.

Sievert, C. and Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In Chuang, J., Green, S., Hearst, M., Heer, J. and Koehn, P. (eds.), *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Baltimore, Maryland: Association for Computational Linguistics, pp. 63–70.

Spooren, P., Brockx, B. and Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Review of Educational Research*, 83(4), 598–642.

Sprague, J. and Massoni, K. (2005). Student evaluations and gendered expectations: What we can't count can hurt us. *Sex Roles*, 53(11–12), 779–793.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503), 755–770.

Wagner, N., Rieger, M. and Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review*, 54, 79–94.

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal*, 12(2), 308–331.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, 12(1), 55–76.