I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13603

Biases in Student Evaluations of Teaching: An American Case Study

Natalia Radchenko

AUGUST 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13603

Biases in Student Evaluations of Teaching: An American Case Study

Natalia Radchenko
American University and IZA

AUGUST 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Biases in Student Evaluations of Teaching: An American Case Study[*]

This work contributes to the literature raising concerns with the use of SET (student teaching evaluation) scores to evaluate teaching effectiveness and to motivate or demotivate faculty tenure and promotion decisions. It shows that the non-deterministic and qualitative nature of the SETs controverts their analysis and interpretation. The evidence of the strong selection of the (un)happiest students into survey participation since the recent switch to the online format with voluntary participation further demonstrates the subjective nature and invalidity of the SETs. The paper also provides empirical evidence of the unidimensionality of the SET answers: various SET items convey uniform content (satisfaction with students' in-class experience) regardless of the questions' specificity. Further, it reinforces empirical evidence that the SET usage introduces multiple biases related to professor, course, and class characteristics and facilitates grade inflation. These biases unbalance the SET scores differently at different tiers of the scores distribution. The results suggest that the biases based on gender and penalty to teaching very weak students are particularly strong. Use of recent and large American data[**] raises the validity and relevance of the findings relating to gender biases induced by the SETs and mainly reported by researchers from European institutions providing large data at the university level.

JEL Classification: I21, I23, J16, J71
Keywords: student teaching evaluation, subjectivity, biases

Corresponding author:
Natalia Radchenko
Department of Economics
American University, Washington DC
4400 Massachusetts Avenue NW
Washington, DC 20016-8029
USA
E-mail: radchenko.nat.au@gmail.com

[*] Grateful acknowledgment to Caryn Bruyere for editorial help.

[**] This research was determined to qualify for exemption under Category 2.(i) of the Code of Federal Regulations Title 45 §46.101(b): Research that only includes interactions involving educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior (including visual or auditory recording). The information obtained is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identiers linked to the subjects.

1 Introduction

This paper aims to inform the continuing discussion regarding the inadequacy of student teaching evaluation scores, SETs, as a measure of professors' teaching efficiency. According the online New World Encyclopedia, the system was pioneered by American psychologist E.T. Guthrie, who served as dean of the graduate school at the University of Washington in the 1940s. Interestingly, "Participation of faculty members was strictly voluntary." In the 1960-1970s, the system spread throughout most North American universities[1] as an instrument of student-teacher communication meant to improve teaching quality. Instead, it progressively became the only tool used to assess professors' teaching competence and an important basis for tenure and promotion decisions, despite concerns regarding multiple biases and the ease of manipulating results to justify personnel decisions.[2]

This evolution from a means of communication to a corporate-oriented assessment of professors' performance engendered a large body of academic literature revealing multiple biases such as gender (the most recent examples are Mengel, Sauermann & Zölitz, 2019; Mitchell & Martin, 2018; Boring, 2016), race (Dee, 2005), and subject matter (Uttl & Smibert, 2017); the literature also shows negative externalities of the dominance of SETs, such as grade inflation (Langbein, 2008; Isely & Singh, 2005; Krautmann & Sander, 1999) and a hindrance to deep learning (Braga, Paccagnella and Pellizzar,2014; Carell, Page & West,2010; Sproule, 2002).

This paper is unique in synthesizing the set of the SET aspects which have been explored separately by the previous studies depending on authors' specific focuses and data sources at hand. First, it contributes to the literature by stating the key statistical problems related to SET usage: the non-deterministic and qualitative nature of SETs, which controverts their analysis and interpretation but is not formally and consistently acknowledged in the related

[1] Becker & Watts (1999) reminded that in 1970s only about 30% of colleges and universities processed SETs, while the SET became omnipresent by 1990s.

[2] Becker (2000): "Less-than-scrupulous administrators and faculty committees may also use them [SETs] because (for the reasons given in the text) they can be dismissed or finessed as needed to achieve desired personnel ends while still mollifying students and giving them a sense of involvement in personnel matters"

literature[3]. The paper also discusses the SET scores from the perspective of behavioral and happiness economics and provides empirical evidence of the unidimensionality of SET answers, which was only acknowledged by Langbein (1994) using data from 1991-1992 and was disregarded by the later literature. Stating the statistical problems is important from the policy perspective as they are ignored by university administrators and faculty committees, who misinterpret SET quantities and use inappropriate statistical measures of teaching effectiveness basaed on SETs.

Next, the paper shows multiple teaching-irrelevant factors underlying the scores, making the SETs biased and incomparable among professors. While many of them have been addressed by the previous literature, this paper reinforces the literature findings by taking advantage of a very large American data set made up of the SET scores and both professor and class characteristics (about 17,000 classes in social sciences involving about 365,000 SET reports and more than 2,000 professors observed over the decade of 2006-2017).

The data in hand allow for updating the findings and proving not only their consistency, but also the coexistence of multiple teaching-irrelevant factors driving the SETs. Use of a recent and large American dataset raises the validity and relevance of findings relating to gender biases induced by SETs and reported so far by researchers from European institutions when using large data at the university level (for example, a Netherlands School of Business and Economics (Mengel, Sauermann & Zölitz, 2019) and a French university specializing in social sciences (Boring, 2016)) .

The paper also provides additional evidence of the SET penalty associated with teaching mandatory and quantitative courses. To the best of our knowledge, the issue of quantitative courses is only addressed by Uttl & Smibert, 2017 and is studied under strong data limitations: only class summary evaluations were available to the authors through the website of a mid-size university which was available to the general public; thus, unlike the present paper, the results reported could not account for various course characteristics or professors'

[3] With the exception of the article of Stark & Freishtat (2014) communicated via the ScienceOpen platform and sketching some points.

heterogeneity. Controlling for professors' fixed effects as well as class size and level enhances the internal validity of the negative impact of teaching quantitative courses and its causal interpretation.

Further, the paper shows a compounding penalty for female professors when teaching quantitative courses. This effect was never reported in the previous literature. Similarly to the data used in this paper, the SET literature builds on social sciences data with few quantitative courses; therefore, the impact can only be efficiently estimated by using a relatively large dataset.

Another novel aspect of the study is the evidence of strong selection of the (un)happiest students into survey participation since the recent switch from the in-class paper-based surveys to online surveys with voluntary participation. This further demonstrates the subjective nature of SETs. It also calls for even stronger caution in the interpretation and usage of SETs given the stronger impact of negative outliers on the mean scores due to the heavy left skewness of the score distribution.

The latter result is yielded by a generalized logistic model accommodating the ordinal scale of the SET scores while allowing for variability in the marginal impacts of various course, class, and professors' characteristics along the SET scores distribution. This approach shows that various biases for which there is evidence differ in the direction and strength of unbalancing the SET scores: biases based on gender or penalties to teaching very weak students might be stronger compared to biases induced by larger class size or moderately lower student performance.

The paper proceeds as follows: the next section describes the data and shows the unidimensionality of the information conveyed by disparate questions of the SET survey; Section 3 explores the observable covariates of the mean SET score, which is customarily adopted by universities as a unique measure of the teaching performance of professors and adds to the empirical evidence of multiple biases reported in the literature; Section 4 elucidates the nature of the SETs data and their means from a statistical standpoint and states the relevant statistical issues; it also re-explores the determinants of the SETs, taking into account the qualitative nature of the data. The Conclusions section summarize the findings.

2 Data

The data come from an American research university hosting several schools, mainly specialized in social sciences while also offering programs in arts, humanities, natural sciences and education. For the sake of uniformity, data comparability and statistical efficiency, this paper focuses on the social sciences data corresponding to the main body of university students, classes and professors.

Several data sources are employed. First is the student teaching evaluation (SET) survey (fall 2006 - fall 2017); second, the registrar data providing information on class size and gender composition; and finally, university catalogs (2009-2018) providing professors' characteristics such as rank and years since appointment. The data and their limitations are described below.

The SET survey is administrated anonymously in the last weeks of each semester but before the final test is administered to the students. Prior to the Spring semester of 2016, the survey was run in class; thus, the respondents were the students who attended class on the day of the survey; the corresponding participation rate ranges from 0.26 to 1, with a mean and median value of 0.9 and skewness of -1.2. Since the Spring of 2016, the survey has been administered online: students complete a questionnaire outside of class, any time within a given time window (typically a couple of weeks before the final test).[4] This part of the survey covers 20% of classes. The response rate decreased in this period: the online survey participation rate ranges from 0.06 to 1, with a mean and median values of 0.64 and skewness of -0.03, implying a more uniform distribution in comparison to the previous period, in which the probability mass was at the right of the rate distribution.

[4] The survey is open to all the students enrolled in the class regardless of their attendance rate. Attendance rate is not available in the data.

The total sample consists of 365,187 individual SET reports from 17,750 social sciences classes offered from Fall 2006 to Fall 2017 and delivered by 2,093 professors. It includes classes from 15 fields of study, such as international service (26.5%), communication (19.5%), public affairs (with 10.5% classes in government studies, 5% in law, and 8% in public administration), economics (6.9%), and business, management, marketing, finance, sociology, anthropology, and psychology making up the remaining quarter of the sample. The working sample based on the complete SET reports providing the answers to all the survey items used in the analysis amounts to 349,136 individual SET reports from 16,694 classes.

Classes

Table 1 shows the distribution of classes across different levels. It also shows distribution of classes by three enrollment categories: small classes of less than 20 students, medium classes of 21-45 students and large classes of more than 45 students. Finally, it reports the class distribution by type: quantitative and elective. The elective nature of the class is reported by the in-class part of the survey and is no longer reported in its more recent online part.

Table 1: Distribution of classes across level (left panel), class size (middle panel), and type (right panel)

Level	%	Size	%	Type	%	%
Undergrad intro or foundational (1)	12.6					
Undergrad intro or foundational (2)	13.8	≤ 20	52.2	Quantitative	2.7	
Undergrad upper-level or advanced (3)	20	20-45	46.5			
Undergrad upper-level or advanced (4)	13	> 45	1.3	Elective		11.9
Advanced Undergrad / MA (5)	9					
MA / PhD (6)	26.8					
PhD (7)	4.5					
PhD (8)	0.3					
Number of classes	16964		16964		16964	11644

Professors and students

The classes in the sample were taught by 2,093 professors. About 40% of professors in each semester are women. About 60% of professors are full-time. Their distribution across rank and years since appointment are shown in Figure 1.

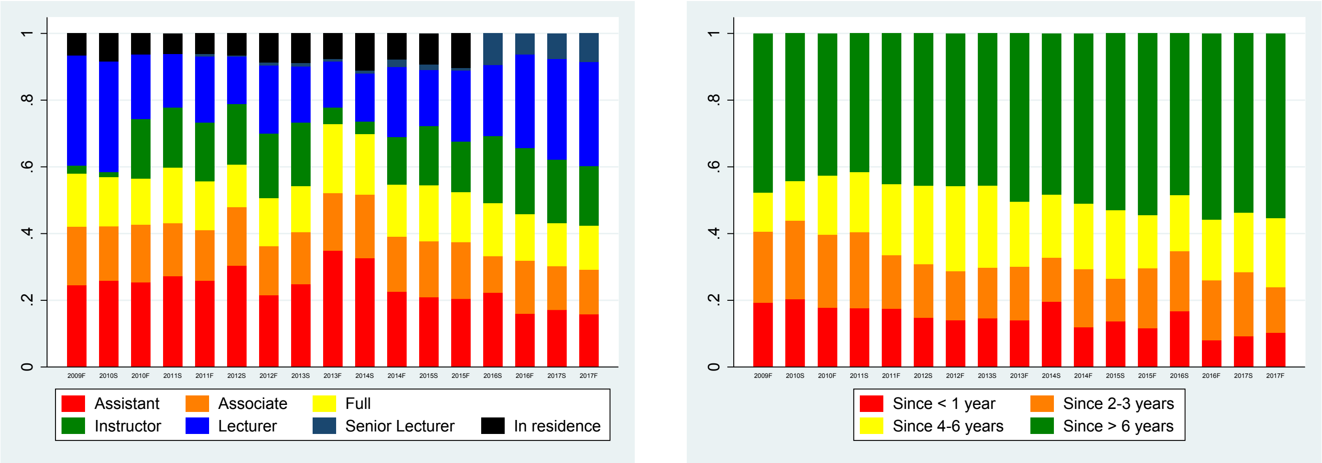

Figure 1: Rank and tenure.

The gender composition of the student body is relatively balanced with a slightly higher share of female students (61% females vs 39% males). This is marginally more unbalanced in classes taught by women (64% female students) than in classes taught by men (58% female students). Students' gender distribution is rather stable over the 2006-2017 period.

Survey

The survey design includes:

- six instructor-related items evaluated on the scale ranging from 1 (the worst) to 7 (the best) score (The instructor used class time productively; The instructor was open to questions and comments; The instructor provided useful feedback on tests, papers, ets; The instructor returned work in a timely manner; The instructor required high levels of performance; On a scale of one to seven, overall the instructor was...);

- five course-related items evaluated on the scale ranging from 1 (the worst) to 7 (the best) score (The learning objectives for this course were clear; Activities/assignments required for the class contributed to meeting the learning objectives; Materials required for this course contributed to meeting the learning objectives; I am satisfied with what I learned in this course; On a scale of one to seven, overall this course was...),
- up to five additional items of the departments' choice,
- and some student items, including the question "What grade do you expect in this course?"

The left graph of Figure 2 displays the distribution of the individual scores corresponding to overall satisfaction with the course ("On a scale of one to seven, overall this course was..."). The mean value and standard deviation are 5.7 and 1.42 respectively. The distribution of the respective class mean scores is shown on the right graph of the same Figure. The standard deviation of the mean score is 0.8.

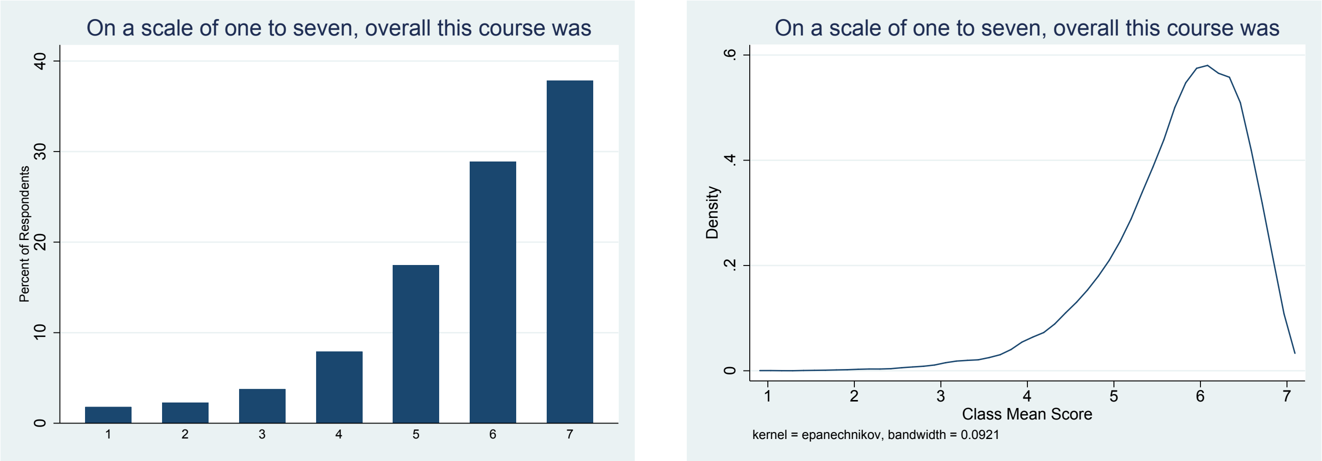

Figure 2: SET score distribution. Left graph: individual reports; right graph: class mean reports.

Table 2 documents the overall distribution of expected grades reported by the students. The distribution of expected grades is independent of the professor's gender. A large majority of students expect to earn either an A or an A- . This is very different from many European institutions unexposed to the North American student teaching evaluation system, which frequently assign the highest grades only to outstanding students or students mastering the subject matter of the course in particular depth. While obviously the SET system is not the only difference among universities in different countries, the empirical analysis in the next sections concurs with the literature relating the SET system to US grade inflation.

Table 2: Students

	Mean % per class	SD
Expected grade		
A	37	22
A-	32	15
B+	15.5	12
B	10.5	11
B-	3	5
C-D-F	2	5
Number of classes	17750	

2.1 Unidimensionality of the SET answers: specificity of a question does not matter

Each question on the survey is supposed to convey information about a unique aspect of the in-class experience of students and to provide multidimensional measures of teaching quality. However, principal components analysis yields only one factor with a high eigenvalue of 7; the eigenvalue of the second vector drops to .79 (the left panel of Table 3). This implies that the first component has the variance of 7, explaining 65% of the total variance of the 11 components corresponding to the 11 standardized questions. Furthermore, the right panel of Table 3 shows that all of the 11 items load positively and nearly equally on the first component.

11 variables converting into one principal component with positive balanced factor loadings implies that the information conveyed by these variables is essentially unidimensional: all the variables are strongly positively correlated and carry the same information - satisfaction with students' in-class experience. This is in line with Langbein (1994), who showed

the unidimensionality of a similar survey using data from 1991-1992. More recent literature also reports about 0.8 pairwise correlations between students' reports on different questions (Ewing, 2012).

Table 3: Principal Components Analysis

Components	Eigenvalue	% of Var	**Principal Component factor loadings**	
			The instructor used class time productively	0.31
1	**7.09**	**65**	The instructor was open to questions and comments	0.27
			The instructor provided useful feedback on tests, papers, ...	0.3
2	.784	7	The instructor returned work in a timely manner	0.29
3	.609	5.5	The instructor required high levels of performance	0.26
4	.538	5	On a scale of one to seven, overall the instructor was...	0.34
5	.460	4	The learning objectives for this course were clear	0.31
6	.384	3.5	Activities/assignments required for the class contributed	0.32
7	.324	3	to meeting the learning objectives	
8	.314	3	Materials required for this course contributed	0.29
9	.218	2	to meeting the learning objectives	
10	.169	1	I am satisfied with what I learned in this course	0.33
11	0.1	1	On a scale of one to seven, overall this course was...	0.33
Number of observations			349,136	

Particularly illustrative is the correlation between the score given to the question on returning work in a timely manner and the overall satisfaction scores. Returning work in a timely manner is, first, under professor's control, and second, is the only objective aspect of the survey. This makes it different from all other items relating to students' feelings about the course and its instructor. With typically uniform within-class policy and professor's control regarding returning graded assignments to the class (i.e. outside of late submissions, students all receive graded assignments on the same day, typically as soon as all grading is complete), the answers to this question should be same within classes, be high-scored and independent of the scores associated with the other items. Yet, Table 4 shows that within-class deviations from the mean class answer vary from -6 to +5 points and that the given score is highly correlated with the score associated with overall course satisfaction:

Table 4: Returning work in a timely manner vs course satisfaction

Returning work in a timely manner	
Range of deviations from the class mean scores	[-6;5]
Equality to the course satisfaction score	50%
1-point difference with the course satisfaction score	33%
Number of observations	349,136

Given the unidemensionality of the survey, the analysis below focuses on one question about the overall satisfaction with the course, which by definition reflects the main information contained in the SET survey: “On a scale of one to seven, overall this course was...”.

3 Mean SET score

The parameter used by university teaching evaluation systems is the unconditional class-mean $\overline{SET}$ score. These unconditional mean scores are frequently compared to the average mean scores of some reference groups, such as department averages, to make decisions relative to tenure and promotion. Leaving aside the qualitative nature of the SET data (discussed in the next section), there are two implicit assumptions underlying such an approach:

1. The unconditional class-mean $\overline{SET}$ score is a deterministic parameter;
2. Variation of the mean score $\overline{SET}$ among professors is due to variant teaching performance.

This section shows that these assumptions are deeply invalid.

3.1 Individual $\overline{SET}$ score is a random outcome

Considering the unconditional class-mean $\overline{SET}$ score as a deterministic parameter implies that the mean score perfectly measures the parameter of interest in the population of interest. This is a deeply flawed approach. If it is overall professor’s performance rather than success in one particular class which is evaluated, a class-level distribution parameter is nothing but a mean score estimate subject to sampling errors. To consider it out of any statistical inference - as it is customarily done - is meaningless from a statistical standpoint. This approach is even more problematic given that the survey participation rate is not 100% and can be as low as 10% (see Section 2); therefore, the parameter is not deterministic, even for a given class.

The sampling errors of the $\overline{SET}$ outcomes depend on the sample size and the distribution of individual SET reports. As shown by the left graph of Figure 2, the SET distribution is strongly skewed to the left. The mass of the distribution is concentrated on the right, over the values of 5-7; the left tail of the distribution corresponds to low satisfaction values of 1-4. The skewness of the distribution implies that the mean value does not describe an average or typical value of the distribution, contrary to what it is designated to represent. The use of the mean value is clearly inappropriate for skewed distributions: it is the most sensitive to the outliers and, most frequently, is below the median unless the distribution is multimodal.

The left skew implies that negative outliers are given a particularly strong weight. Consider a specific class of 13 students drawn from the actual sample. The satisfaction reports delivered by 7 out of 13 students were $\{1, 5, 6, 6, 7, 7, 7\}$. The mean score is 5.57, which is 0.5 below the same-term department average of 6.08 and makes the professor's "performance" appear to lag behind that of his colleagues. The presence of one negative outlier, reporting the value of 1, appears to depress the mean substantially. This example obviously illustrates the unbalanced weight given to one dissatisfied student: given the left-skewed distribution of the scores, only a dissatisfied student can be an outlier; a negative outlier contribution is never counterbalanced by the most highly-satisfied student, whose happiness is limited by the highest possible value M, with $M = 7$ in this specific case study. The loss $\frac{\Delta_i}{N}$ from a record of student i reporting Δ_i points below the highest SET value, depends on the class size, N:

$$\overline{SET} = M - \sum_{i=1}^{N} \frac{\Delta_i}{N}$$

For example, a value of 1 reduces the maximum possible mean score of 7 by $\frac{6}{N}$, where N denotes the class size. In the given example, it reduces the highest-possible SET by 0.86, which greatly exceeds the lag of 0.5 behind the reference value of 6.08. In fact, in this case, dropping the outlier would have yielded a mean score of 6.33: the presence of one dissatisfied student in this case results in a professor who otherwise would have earned a result exceeding expectations instead earning a score appearing to underperform. Only by earning perfect satisfaction scores from all six other students could the professor attain a score at or above the department average of 6.08.

3.2 $\overline{SET}$ score is multifaceted

The variation of the mean $\overline{SET}$ score across different classes is assumed to be due to the variant teaching performance of different professors. The analysis below invalidates this assumption, showing that the source of variation in the mean score is multifaceted. It reports the determinants of the mean SET variation given the data available.

Using the mean SET scores as a dependent variable implies the grouped data regression, where grouping is done at the class level, c:

$$\overline{SET}_{prof,class,field,sem,t} = \beta_0 + \mathbf{X}_{prof,class,sem}\boldsymbol{\beta} + \gamma_{prof} + \delta_{field} + \sigma_{sem} + \lambda Online_t + \epsilon_{prof,class,sem,t} \quad (1)$$

where $\mathbf{X}_{prof,class,sem}$ is a vector of class-level observable characteristics of the class (level, enrollment, semester, class type) and students (gender composition and distribution of the expected grades); $\boldsymbol{\beta}$ is the vector of the corresponding coefficients; γ_{prof} are individual professors' fixed effects; δ_{field} are the study field fixed effects, allowing us to control for distinct student bodies and heterogeneity of schools/departments providing classes in different fields; σ_{sem} is the semester fixed effects allowing for overall correlations across spring, summer or fall semesters. $Online_t$ is an indicator of computer-based out-of-class survey ($Online = 1$ for the period t of 2016-2017 and $Online = 0$ for the period t of 2006-2015[5]), λ is the corresponding online parameter, and $\epsilon_{prof,class,field,sem,t}$ is a vector of unobservable $\overline{SET}$ determinants.

Several series of model (1) estimates are reported in Table 5-6: the first column corresponds to the full sample of 2006-2017; the second one reports the estimates based on 2006-2015, which allows us to see the effect of the elective class type[6]; the last column shows the estimates based on 2009-2017, allowing augmentation of the model with professors' gender, rank and work experience in the university[7]. Unlike model (1), the latter analysis is run without the professors' fixed effects because of the invariability of the professor's own characteristics[8]:

$$\overline{SET}_{prof,class,field,sem,t} = \beta_0 + \mathbf{X}_{prof,class,sem}\boldsymbol{\beta} + \mathbf{Y}_{prof}\boldsymbol{\alpha} + \delta_{field} + \sigma_{sem} + \lambda Online_t + \epsilon_{prof,class,field,sem,t} \tag{2}$$

where $\mathbf{Y}_{prof}$ and $\boldsymbol{\alpha}$ are the variable and coefficient vectors associated with the professors' characteristics.

Empirical findings

The results reported in Table 5-6 show that the SET scores are driven by multiple factors unrelated to teaching. First, the average SET scores differ with observable class and course characteristics. Specifically, the mean scores are on average 0.12 points lower in graduate-level courses, in particular those for master's-level (MA) programs. MA students are more frequently involved in the labor market or soon will be. Consequently, they are more concerned with the grades as signals of their performance for employers and are more demanding relative to the match between the course level or content and their individual background

[5] The parsimonious model presented does not include more detailed year fixed effects since they are found weak in magnitude, statistically insignificant and uncorrelated to other regressors when included.

[6] As mentioned above, the elective type is not reported by the more recent, online part of the survey.

[7] As mentioned above, the rank and years since appointment are only available from the university catalogs since 2009 unlike the SET data recorded since 2006.

[8] The only exception is the professor's rank; yet, it has very low variability within the observational window.

and needs.

Not only the level of the course but also its type matters in student teaching evaluations. Unsurprisingly, teaching elective courses is rewarding: the mean SET score is on average 0.08 higher in elective classes. On the other hand, as reported by previous research, teaching a quantitative course to social sciences students might be penalizing (Uttl & Smibert, 2017, Langbein, 1994).

According to model (1) estimates, this is particularly true for female professors: while not very precise (relatively weak efficiency of the estimate is not surprising given the low number of the corresponding observations), the coefficient associated with the quantitative type of the course when taught by women is negative and implies up to a 0.16 decrease in SET score as displayed in column 2 of panel I (Table 5), on average. Using French data, Boring (2016) shows that gender stereotypes strongly drive student teaching evaluations. She finds that students reward or penalize male and female professors on teaching dimensions typically associated with gender. They associate male professors with competent and authoritative personality, penalizing female professors on these grounds but rewarding them on the dimensions relating to warm and nurturing behavior. The finding of a negative impact of teaching a quantitative course is likely to be of the same nature: social science students are naturally less inclined to technical subject matters and frequently have weak technical backgrounds which makes the quantitative courses dreaded and difficult for many of them; to dismiss their difficulties, some students might charge professors with incompetence and more easily do so with regards to female professors.

The estimates issued by model (2) (reported by panel III of Tables 5-6) do not yield the same gender effect relative to the quantitative courses, but they do show a significant overall gender gap of 0.17 points in favor of men (see Table 6, column 1 of panel III). Remember, this model integrates some invariant observable professors' characteristics (including gender) at the cost of omitted professor fixed effects and a sample reduced by about 25%. It allows us to see that the mean scores of new appointees are the lowest, keeping other observables (including the professor's rank) equal (Table 6). After the first year of teaching, the mean score increases on average by 0.2 points; it reaches its maximum within the first three years of teaching in a given institution, resulting in a 0.25 point gap with the first year of teaching and remains constant after that. Note that these effects are found while controlling for professors' rank. This means that it is not necessarily additional teaching experience per se which allows teachers to gain higher SET scores, but adjustments to the student body and eventually new courses in a given institution.

Class size and composition are also strong determinants of the SET (Table 5). The size of the class has a progressively negative impact on the mean SET scores: teaching "medium" (21-45 students) vs "small" classes of 20 or less students yields an average loss of 0.16 points; teaching large classes of more than 45 students depresses the mean score twice as strongly (by about 0.3 points).

The gender composition effects support the findings evidenced by French data and showing that first, a gender match between the student and professor raises the SET score (Boring, 2016) and second, that female students are less generous in ranking professors as compared to male students. Indeed, as reported by Table (5), the higher the percentage of female students, the lower on average the mean SET score: teaching a female class vs a male class would depress the mean score by up to 0.4 points for male professors (column 1 of panel I, Table 5). However, the negative effect is about twice as weak for female professors as shown by column 2 compared to column 1 of panels I, Table 5, indicating a positive effect of student-professor gender match.

A classical result documented by the previous research is a strong progressive impact of the grades expected by students by the end of semester. Table 6 show that the mean SET scores are elastic relative to an increasing percentage of students earning lower grades: the lower the grade, the lower the score. Expecting an A-, B+ , B or B- instead of an A in the class decreases the class mean SET score by 0.1, 0.35, 0.9 and 1.5 respectively; any class expectation of a grade below a B- decreases the mean score by 1.6.

Intuitively, the results imply that students expecting higher(lower) grades, reward(punish) professors by higher(lower) evaluations. The interpretation might be clouded by revers causality: universities award professors with tenure and promotion for higher SET scores; professors might therefore inflate students' grades to "buy" high SETs. Both ways go along the lines of Langbein (2008) who conceptualizes the grades as the currency binding the values of students, faculty and administrators in parallel with prices, which bind consumers, firm employees, and managers. Note however, that the results come out of the regression with professors' fixed effects which indirectly control for professors' unobserved characteristics and practices potentially impacting the SET scores. Those include faculty grading policies and teaching approaches. They therefore also reduce a potential bias coming from another source of exogeneity which could be simultaneous rise of grades and SETs in response to better instructions.

Should the relationship between the SETs and expected grades be driven by the instruction quality, it would not be impacted by professors's characteristics irrelevant to teaching such as professor's gender. However, the results yield the strong and strikingly robust gender gap in grade elasticity of the mean SET score: according to model (1) estimates, the effect is about 0.1-0.5 points stronger if the instructor is a woman (column 2 of panels I and II, Table 6). The gender gap is particularly strong in the case of grades below B- and increases to 1 point. These results are in line with Boring (2016), Sinclair & Kunda (2000), and the gender gap reported above implying that students apply different standards to male and female professors: weakly-performing students are more likely to attribute their low performance to the professor rather than to themselves if the professor is a woman.

Finally, the estimation yields a positive summer semester effect of 0.14, which is likely to be related to the more elective nature of summer classes as well as more relaxing and joyful time period. Students may also be less distracted by extracurricular activities or the need to balance the demands of several simultaneous courses, as opposed to one or two.

Table 5: Mean SET scores

	I. 2006-2017, all		II. 2006-2015, all		III. 2009-2017, Full-time	
	Men	Women vs Men	Men	Women vs Men	Men	Women vs Men
Class						
Intro Undergrad (2)	0.040*		0.036		0.008	
	(0.023)		(0.029)		(0.030)	
Upper Undergrad (3)	0.021		0.031		-0.064**	
	(0.021)		(0.026)		(0.029)	
Upper Undergrad (4)	0.019		0.041		-0.052	
	(0.024)		(0.030)		(0.032)	
Adv. Undergrad/MA (5)	0.111***		0.103***		0.080**	
	(0.027)		(0.032)		(0.039)	
MA/PhD (6)	-0.120***		-0.115***		-0.183***	
	(0.023)		(0.029)		(0.030)	
PhD (7)	-0.059*		-0.086**		-0.152***	
	(0.033)		(0.041)		(0.045)	
PhD (8)	-0.016		0.022		0.065	
	(0.086)		(0.116)		(0.102)	
Elective			0.078***	-0.031		
			(0.025)	(0.039)		
Quantitative	0.001	-0.163*	-0.029	-0.097	-0.060	-0.083
	(0.051)	(0.089)	(0.065)	(0.111)	(0.053)	(0.103)
Enrollment $<$ 20						
Enrollment 20-45	-0.164***	-0.055**	-0.166***	-0.059**	-0.109***	-0.063*
	(0.015)	(0.022)	(0.019)	(0.028)	(0.023)	(0.032)
Enrollment$>$45	-0.299***	-0.117	-0.214***	-0.186	-0.143*	-0.088
	(0.060)	(0.090)	(0.078)	(0.115)	(0.086)	(0.129)
Female students fraction	-0.391***	0.220***	-0.335***	0.145*	-0.323***	0.287***
	(0.045)	(0.070)	(0.057)	(0.087)	(0.066)	(0.090)
Fall semester	0.019**		-0.003		0.025	
	(0.010)		(0.012)		(0.016)	
Summer semester	0.135***		0.144***		0.143***	
	(0.023)		(0.030)		(0.042)	
Online survey	-0.000		0.000		0.080***	
	(0.014)		(.)		(0.019)	
Constant	6.652***	6.652***	6.621***	6.621***	6.136***	6.136***
	(0.134)	(0.134)	(0.191)	(0.191)	(0.069)	(0.069)
Study Field	Yes	Yes	Yes	Yes	Yes	Yes
Observations	16964	16964	11025	11025	8136	8136

Standard errors (in parentheses) are clustered at the professor level
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Mean SET scores, continued

	I. 2006-2017, all		II. 2006-2015, all		III. 2009-2017, Full-time	
	Men	Women vs Men	Men	Women vs Men	Men	Women vs Men
A (Expected grades, frac)						
A-	-0.104**	-0.151**	-0.053	-0.252***	0.004	0.033
	(0.048)	(0.074)	(0.062)	(0.095)	(0.074)	(0.112)
B+	-0.352***	-0.158*	-0.258***	-0.222*	-0.596***	0.393***
	(0.060)	(0.093)	(0.078)	(0.120)	(0.092)	(0.138)
B	-0.896***	-0.123	-0.854***	-0.128	-0.969***	-0.203
	(0.070)	(0.108)	(0.090)	(0.140)	(0.108)	(0.163)
B-	-1.487***	-0.131	-1.207***	-0.497*	-1.592***	-0.225
	(0.134)	(0.208)	(0.181)	(0.276)	(0.219)	(0.326)
C-D-F	-1.583***	-0.631***	-1.116***	-1.119***	-2.071***	-0.637*
	(0.146)	(0.227)	(0.191)	(0.292)	(0.228)	(0.350)
Professor						
Female					-0.167**	
					(0.081)	
Appointed since $<$ 1 year						
Appointed since 2-3 years					0.196***	
					(0.028)	
Appointed since 4-6 years					0.253***	
					(0.028)	
Appointed since $>$ 6 years					0.239***	
					(0.029)	
Assistant						
Associate					-0.124***	
					(0.026)	
Full					-0.077***	
					(0.027)	
Instructor					0.067*	
					(0.038)	
Lecturer					-0.092***	
					(0.026)	
Senior Lecturer					-0.242***	
					(0.044)	
In Residence					-0.045	
					(0.035)	
Professor FE	Yes		Yes		No	
Within R-sq	0.110	0.110	0.108	0.108		
Observations	16964	16964	11025	11025	8136	8136

Standard errors (in parentheses) are clustered at the professor level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4 SET Score is an Essentially Subjective Ordinal Measure

The previous section focused on the class-mean $\overline{SET}$ score as a numerical measure conventionally adopted by universities; it restated the non-deterministic and multidimensional nature of the score, explaining the inadequacy of the current practice of teaching evaluation. This section states another problem related to the usage of SET, which is its qualitative nature. The section then reconsiders multiple determinants of the score, taking into account its intrinsic nature.

4.1 Subjective data limitations in the SET context

The data reported by the SETs survey are what is typically qualified as “subjective data” in social sciences. This kind of data is very cautiously used in happiness economics and inequality studies. There are a number of very standard concerns limiting this literature (Di Tella & MacCulloch, 2006, van Praag, 2007):

a) Interpretation of the answers is questionable: what do the individuals have in mind when reporting their satisfaction?

b) From a statistical standpoint, subjective data are “qualitative data”, not “quantitative”. The different numbers assigned to different satisfaction levels are no more than labels; they do not represent any quantity or magnitude. The ranking data are ordinal at best: assuming that individuals assigning the same scale value, experience identical levels of satisfaction, the different levels can be ordered; however, numerical distances between different levels are undefined and do not make sense. Numerical parameters such as unconditional or conditional mean scores are intrinsic to cardinal measures; they are meaningless when applied to qualitative data.

c) The assumption that different individuals experience the same levels of satisfaction when assigning the same scale values is not warranted either: Even when referring to the same values, individuals do not necessarily provide comparable answers because they might have different scales in mind when responding; individual scales might be driven by latent psychological factors and individual-specific backgrounds. For example, female respondents typically give lower responses as compared to male respondents (Bertrand, 2011, Stevenson & Wolfers, 2009) regardless of the issue of satisfaction individuals are questioned about (life, income, etc.). While there might be weaker women's endowments leading to objectively lower satisfaction in some contexts, the gaps always go the same direction and are never entirely explained.

The scales are impacted by individual psychological factors (pessimism, optimism), emotion-related aspects/the context of the moment when a survey is run, and mood at the time of judgment. For example, a large body of empirical literature in happiness economics shows that one's subjective welfare and satisfaction rises with nice meteorological conditions (e.g. Connolly, 2013; Schwarz & Clore, 1983). Social psychology experiments also show that the respondents' momentary affective state impacts their current judgments of any kind. For instance, Schwarz & Clore (1983) find that individuals in "unpleasant affective states" are particularly likely to attribute their moods to external sources and relate their states to any satisfaction judgments of the moment; individuals experiencing positive feelings are less likely to link their states to information processing and judgments.

The concerns listed in a),b), and c) pertain to the student evaluations of teaching:

a) The interpretation and validity of the students' answers is a great issue in the literature: what are the determinants of students' satisfaction with a course versus the professors' abilities, which administrators seek to evaluate using the SETs? If the latter is teaching efficacy, the literature shows that the two sets of parameters are not the same since the learning outcomes are far from the students' only value and objective: as shown above and in the previous literature, students value grades, confusing them with the value of education; many students also value entertaining aspects of time spent in class, their enjoyment (using a randomized experience, Hessler et al. (2018) show that providing chocolate cookies in class has a positive impact on the SETs), and effortless learning processes (Boring, Ottoboni and

Stark, 2016; Braga et al., 2014; Stark & Freishtat, 2014; Langbein, 2008).

Further, students are not necessarily able to evaluate the professors' competence due to inexperience and insufficient qualifications as compared to the professionals they are invited to judge (Hornstein, 2017; Stark & Freishtat, 2014).

b) To see the nonsense of the cardinal approach, reconsider an example from section 3.1. Assume that the negative outlier reported 4 (still a rather "negative" score) instead of 1. The class sample would then be $\{4, 5, 6, 6, 7, 7, 7\}$, with the new mean estimate of 6 making the professor's "performance" very close to the department average, thus a satisfactory performance. The SET system implies that distrubutions $\{1, 5, 6, 6, 7, 7, 7\}$ and $\{4, 5, 6, 6, 7, 7, 7\}$ are very different in terms of teaching productivity, which is obviously senseless: would the reported value of 4 mean that the negative outlier is 4 times happier than when he reported 1? That he learned 4 times more or 4 times better? That the professor is 4 times more efficient? Clearly, this is not the case.

c) As shown above and in Boring (2016), female students tend to give lower teaching evaluations as compared to male students. The use of the scale might differ not only between the gender groups but more broadly among different types of students: students with different backgrounds, learning approaches, and levels of maturity and responsibility are likely to interpret and use the same values of the scale differently. An example is the heterogeneity of within-class reports on "Returning work in a timely manner" (section 2.1). Inconsistency of the answers regarding a presumably objective aspect of professors' work demonstrates irresponsibility of some students; irresponsibility which is easily elicited by the survey's anonymity.

Psychological aspects of the SET survey outcomes as reported in the literature show that SET scores depend on the day and time of the survey (Boring, 2016, Braga et al., 2014). For example, Braga et al. (2014) find that the scores are positively correlated with outside temperature and negatively with rainy days. It illustrate that, as with any other subjective data, the SET rankings are subjected to irrelevant environmental and psychological conditions and shocks, negative ones in particular. The positive summer semester effect shown above supports this point. Thus, students experiencing identical levels of satisfaction may assign different scale values to the course, and students assigning the same scale value may experience differing levels of satisfaction

The next subsection exploits the ordinal nature of the SET data to provide additional empirical evidence of the biases discussed in the previous sections.

4.2 Biases detailed by the ordinal approach

The Generalized logistic model accommodates the ordinal scale of the SET scores. This approach has two advantages. First, unlike the expected value models, the generalized logistic model allows the exploration of the SET determinants along the SET scores distribution; second, it takes into account the ordinal nature of the SET scores by respecting different categories of the scores and the ordered relationship between different score levels while remaining insensitive to the numeric distances between different values:

$$\begin{aligned} P(SET=1) &= P(SET<2) = 1-\mathit{\Lambda}(\alpha_2-\beta_2 X) \\ P(SET=j) &= P(SET<j+1)-P(SET<j) = \mathit{\Lambda}(\alpha_j-\beta_j X)-\mathit{\Lambda}(\alpha_{j+1}-\beta_{j+1} X) \\ P(SET=7) &= 1-P(SET<7) = \mathit{\Lambda}(\alpha_7-\beta_7 X) \end{aligned}$$

with $\mathit{\Lambda}(z) = \frac{exp(z)}{1+exp(z)}$ representing the logistic distribution and $j = 2, ..., 6$.

Note that unlike the standard ordered model, β_j is alternative-specific in this generalized framework. Therefore, the odds ratios describing the probability of an unsatisfied student versus the probability of a satisfied student depend on the level of dissatisfaction, j: $\frac{P(Unsatisfied)}{P(Satisfied)} = \frac{P(SET\leq j)}{P(SET>j)} = exp(\beta_j) \neq exp(\beta)$.

The model can be estimated by a series of binary logistic regressions

$$P(SET<j) = 1-\mathit{\Lambda}(\alpha_j-\beta_j X)$$

with $j = 2, ..., 7$.

A positive β_j yields a higher probability of a lower SET score, in other words, a stronger level of student dissatisfaction.[9]

The estimates are reported in Table 7 (see Appendix). In light of section 2.1 showing the unidemensionality of the survey, the analysis is run using the question about the overall satisfaction with the course ("On a scale of one to seven, overall this course was...") which by its definition reflects the main content conveyed by the SET survey reports.

Figures 4 and 5 illustrate the results in terms of the odds ratios along with the corresponding confidence intervals. An odds ratio can be interpreted as a number of unsatisfied vs a number of satisfied students. The figures show the impacts of regressors on the odds ratios while defining dissatisfaction in different ways: from the narrowest (U_1 corresponds to $SET = 1$) to the broadest (U_6 corresponds to $SET \leq 6$).

[9] A limitation of this model is that individual heterogeneity is not accounted for - not only at students' level, as in the previous section - but also at the faculty level. This is a technical limitation related to the inappropriateness of both fixed and random effect models. Specifically, the random effects logit model is based on the assumption of normal distribution of the individual effects; given the distribution of the SET scores, such an assumption is invalid. The fixed effects logit model is not feasible because it operates conditional on time variance of the dependent variable; however, for many instructors, receiving for example scores of 4-7 vs 1-3 is time invariant.

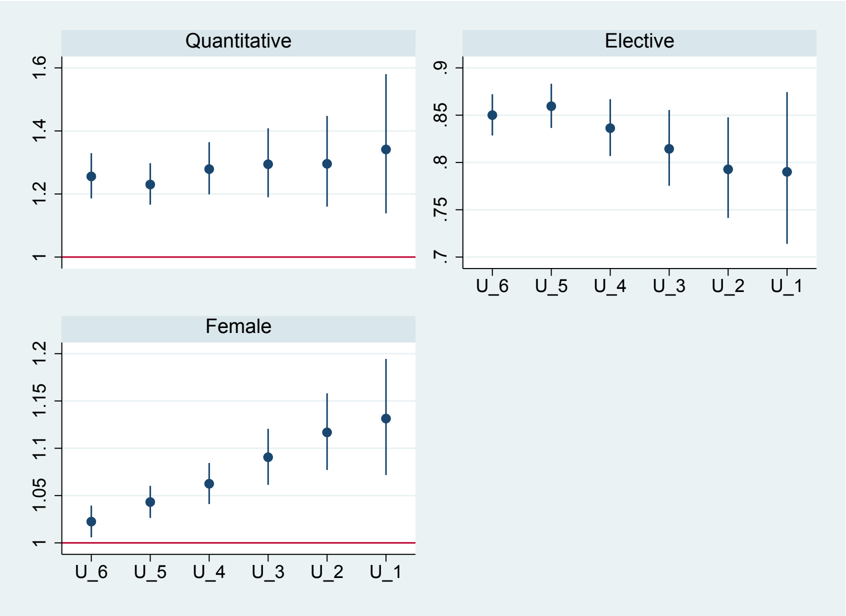

Figure 3: Impacts of the class type and size on the SET odds ratios at different levels of satisfaction reported (U6: 1-6 vs 7, U5: 1-5 vs 6-7,..., U1: 1 vs 2-7)

In line with the previous results, teaching technical and/or larger classes (above 20 students) is penalizing in terms of the SET scores (Figures 3 and 4). Teaching classes of more than 45 students particularly reduces the odds of receiving the highest SET scores. On the other hand, the results conform to the previous finding showing that students find elective classes to be considerably more pleasant (Figure 3): the estimates based on the earlier paper-based part of the survey of 2006-2015 show that the odds of dissatisfaction in elective classes are less than 1, implying higher numbers of satisfied versus unsatisfied students (Figure 3). Moreover, the higher the degree of dissatisfaction in question, the lower the relative number of dissatisfied students as compared to non-elective classes. The effects of other covariates are robust: the estimation based on the reduced sample of 2006-2015 yields very similar results to those based on the whole sample.[10]

The bottom graph of Figure 3 implies a dynamic gender bias of the SET reports. Female professors are more likely to get lower scores, and the bias increases with the level of dissatisfaction: the odds ratio at $j = 6$ (U_6: $SET \leq 6$ vs $SET = 7$) is only 1% higher for female

[10] Not reported but available on request.

professors as compared to male ones. This implies that the odds of receiving 7 relative to any other score is about 1% lower for women as compared to men. The gap strengthens along the dissatisfaction scale, with a maximum of 10% for the odds of receiving 1 relative to above 1 (U_1: $SET = 1$ vs $SET > 1$).

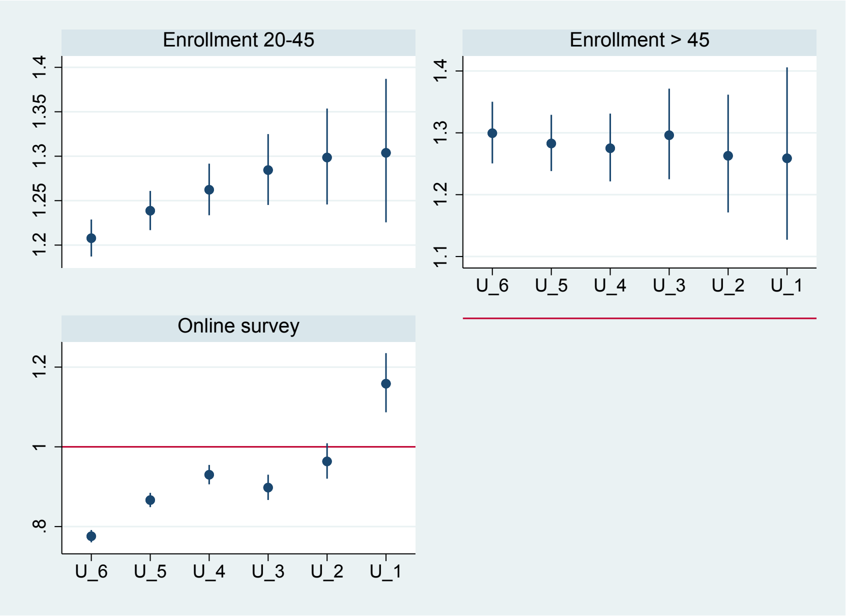

Figure 4: Impacts on odds ratios at different levels of satisfaction reported (U6: 1-6 vs 7, U5: 1-5 vs 6-7,..., U1: 1 vs 2-7)

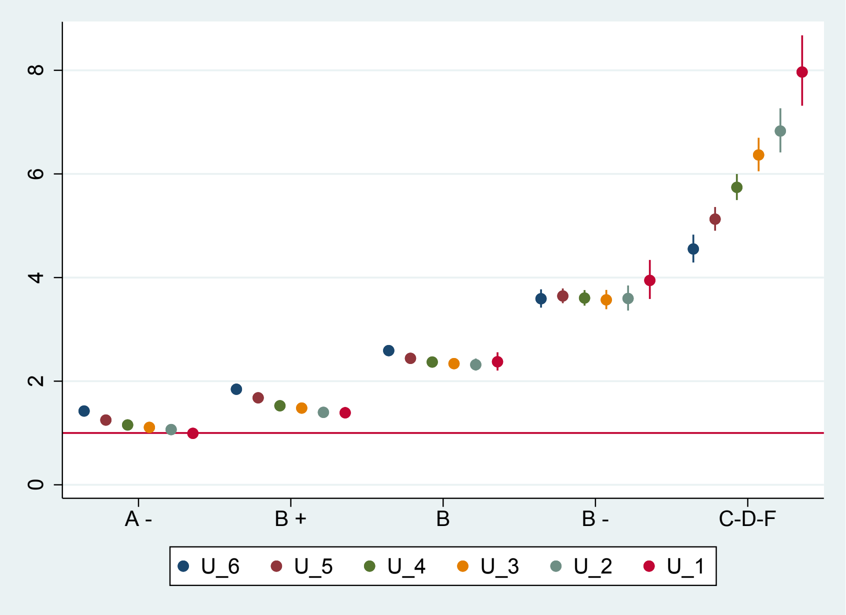

Figure 5: Impacts of expected grades on the SET odds ratios at different levels of satisfaction reported (U6: 1-6 vs 7, U5: 1-5 vs 6-7,..., U1: 1 vs 2-7)

Figure 5 details the impact of the grade the student expects to earn on the odds ratios of unsatisfied to satisfied students. Expecting an A- rather than an A halves the odds of selecting the highest scores, 6-7. The lower the expected grade, the stronger and broader its effect on the odds ratios. Expecting a C, D, or F motivates very strong student dissatisfaction: the probability of selecting the lowest SET score relative to the score above 1 peaks to 8.

The ordinal approach shows that the biases evidenced in the previous sections unbalance the SET scores differently at different tiers of the score distribution: larger classes and lower student performance operate at the upper tier of the score distribution and reduce the likelihood of the highest scores relative to middle values; the biases based on gender or very poor performance strengthen at the lower tier of the SET distribution. The latter risks a stronger impact on the mean scores of the professors concerned because of the strong skewness of the score distribution to the left and consequently the stronger weight of negative outliers.

Finally, Figure 4 shows that the switch from the in-class paper-based survey to the online survey with voluntary participation increased the probability of extreme values particularly strongly: both the probability of 7 relative to below 7 (the left lower corner on the graph) and 1 relative to above 1 (the right upper corner) increased by roughly 20%. This implies that voluntary out-of-class computer-based survey yields the most reports from the most excited/angry students. The fact that students have higher willingness to report extreme values evidences that there is an emotional determinant of SET scores.

Conclusions

There are recent precedents set, for example by the University of Southern California ("Teaching Eval Shake-Up") and Ryerson University of Canada ("Arbitrating the Use of Student Evaluations of Teaching"), which have returned to no or very limited and only formative usage of the SETs with no use towards tenure and promotion decisions. The empirical evidence and discussion presented in this paper show that this new North American trend of renouncing SETs is the right way to avoid inadequate or arbitrary statistical manipulation of scores, discrimination against some categories of professors, and further grade inflation. This paper documents statistical problems associated with the usage of the SET scores to measure teaching performance and empirically analyzes multiple biases:

- Sampling errors and the distributional effects of the individual scores of student teaching evaluation;
- The qualitative and subjective nature of the scores, invalidating their current interpretation and any numeric measurements such as the mean value;
- Unidimensionality of the SET survey, implying the uniform content (satisfaction with students' in-class experience) conveyed by the survey answers regardless of the questions' specificity;
- Multiple determinants of the SET scores irrelevant to teaching ability but penalizing

some categories of the professors.

The results imply that the SET scores are discriminatory and biased on the professor's gender, class size, expected grades, and the nature of the course. Specifically, it is penalizing for faculty in social sciences to teach quantitative, non-elective or larger class under the SET system. Next, while this study is not designed to identify the structural impact of the expected grades on the SETs, it evidences importance of the value that students place on their grades when reporting the SETs. Strong positive impacts of the expected grades imply that the system encourages grade inflation and a consumer orientation of students relative to educational outcomes. Furthermore, it disregards the different standards applied by students to male and female professors due to gender stereotypes. Female professors are particularly penalized for teaching more technical content or assigning lower grades. These biases unbalance the SET scores differently at different tiers of the scores distribution. The switch from the in-class paper-based survey to the online survey with voluntary participation risks to bias further the SET by the selection of the unhappiest and happiest students into survey participation.

Overall, these results imply that the mean SET scores are not a useful basis of comparison as a means to provide measurable differences among professors and across classes. Their comparison to an average value of a reference group such as a department is accordingly not a useful measure of professors' performance or teaching ability.

APPENDIX

Table 7: Generalized Logistic Regression: Unsatisfied vs Satisfied Students

	1-6 vs 7	1-5 vs 6-7	1-4 vs 5-7	1-3 vs 4-7	1-2 vs 3-7	1 vs 2-7
Female professor	0.013*	0.035***	0.051***	0.073***	0.086***	0.095***
	(0.008)	(0.008)	(0.010)	(0.013)	(0.017)	(0.025)
Quantitative	0.198***	0.174***	0.210***	0.206***	0.221***	0.215***
	(0.026)	(0.025)	(0.030)	(0.040)	(0.051)	(0.075)
Enrollment 20-45	0.189***	0.214***	0.233***	0.250***	0.261***	0.265***
	(0.009)	(0.009)	(0.012)	(0.016)	(0.021)	(0.032)
Enrollment $>$ 45	0.262***	0.249***	0.243***	0.259***	0.233***	0.230***
	(0.020)	(0.018)	(0.022)	(0.029)	(0.038)	(0.056)
A-	0.353***	0.223***	0.144***	0.103***	0.064***	-0.007
	(0.009)	(0.009)	(0.012)	(0.017)	(0.023)	(0.036)
B+	0.612***	0.518***	0.422***	0.393***	0.335***	0.330***
	(0.011)	(0.011)	(0.014)	(0.019)	(0.026)	(0.039)
B	0.952***	0.892***	0.863***	0.850***	0.841***	0.865***
	(0.013)	(0.012)	(0.015)	(0.019)	(0.026)	(0.038)
B-	1.279***	1.293***	1.282***	1.273***	1.280***	1.372***
	(0.025)	(0.020)	(0.021)	(0.027)	(0.034)	(0.049)
C-D-F	1.515***	1.635***	1.748***	1.851***	1.921***	2.075***
	(0.030)	(0.023)	(0.022)	(0.026)	(0.032)	(0.043)
Spring Semester	-0.006	-0.005	0.005	-0.002	-0.010	-0.057**
	(0.007)	(0.007)	(0.009)	(0.013)	(0.017)	(0.025)
Summer Semester	-0.171***	-0.230***	-0.280***	-0.305***	-0.327***	-0.458***
	(0.019)	(0.021)	(0.028)	(0.039)	(0.052)	(0.081)
Online survey	-0.254***	-0.143***	-0.073***	-0.108***	-0.037	0.147***
	(0.010)	(0.011)	(0.013)	(0.018)	(0.024)	(0.033)
Constant	0.075	-1.331***	-2.394***	-3.143***	-3.808***	-4.676***
	(0.090)	(0.100)	(0.135)	(0.182)	(0.243)	(0.360)
Class level (2-8)	Yes	Yes	Yes	Yes	Yes	Yes
Field	Yes	Yes	Yes	Yes	Yes	Yes
N	349136	349136	355001	349136	349136	349136

Standard errors (in parentheses) are clustered at the professor level
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

References

Becker, W. (2000). Teaching economics in the 21st century. *Journal of Economic Perspectives*, 14, 109–120.

Becker, W. & Watts, M. (1999). How departments of economics evaluate teaching. *The State of Economic Education*, 89(2), 344–349.

Bertrand, M. (2011). New perspectives on gender. *Handbook of Labor Economics, Elsevier.*

Boring, A. (2016). Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145, 27–41.

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

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students evaluations of professors. *Economics of Education Review*, 41, 71–88.

Carrell, S., Page, M., & West, J. (2010). Sex and science: how professor gender perpetuates the gender gap. *Quarterly Journal of Economics*, 125(3), 1101–1144.

Connolly, M. (2013). Some like it mild and not too wet: The influence of weather on subjective well-being. *Journal of Happiness Studies*, 14, 457–473.

Dee, T. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, 95(2), 158–165.

Di Tella, R. & MacCulloch, R. (2006). Some uses of happiness data in economics. *Journal of Economic Perspectives*, 20(1), 25–46.

Ewing, A. (2012). Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review*, 31(1), 141–154.

Hessler, M., Popping, D., Hollstein, H., Ohlenburg, H., Arnemann, P., Massoth, C., Seidel, L., Zarbock, A., & Wenk, M. (2018). Availability of cookies during an academic course session affects evaluation of teaching. *Medical Education*, 52, 1064–1072.

Hornstein, H. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, 4(1), http://dx.doi.org/10.1080/2331186X.2017.1304016.

Isely, P. & Singh, H. (2005). Do higher grades lead to favorable student evaluations? *Economics of Education Review*, 36(1), 29–42.

Krautmann, A. & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review*, 18, 59–63.

Langbein, L. (1994). The validity of student evaluations of teaching. *Political Science and Politics*, 27(3), 545–552.

Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review*, 27, 417–428.

Mengel, F., Sauermann, J., & U., Z. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, 17(2).

Mitchell, K. & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics*, 51(3), 648–652.

Schwarz, N. & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45, 513–523.

Sinclair, L. & Kunda, Z. (2000). Motivated stereotyping of women: she's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin*, 26(11), 1329–1342.

Sproule, R. (2002). The underdetermination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review*, 21(3), 287–294.

Stark, P. & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research*.

Stevenson, B. & Wolfers, J. (2009). The paradox of declining female happiness. *American Economic Journal: Economic Policy*, 1(2), 190–225.

Uttl, B. & Smibert, D. (2017). Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ*, (pp. DOI 10.7717/peerj.3299).

van Praag, B. (2007). Perspectives from the happiness literature and the role of new instruments for policy analysis. *CESifo Economic Studies*, 53(1), 42–68.