Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness

Anne Boring, Kellie Ottoboni, Philip B. Stark

Draft September 23, 2015

The truth will set you free, but first it will piss you off.

Gloria Steinem

Abstract

We examine whether student evaluations of teaching (SET) primarily measure teaching effectiveness, using nonparametric tests applied to two datasets. For the first dataset, 23,001 SET of 379 instructors by 4,423 students over five years in six different mandatory first year courses arising from a natural experiment at a French university, we study relationships among SET and the genders of students and instructors, final exam scores, and students' grade expectations. For the second dataset, 43 SET for 4 sections of an online course arising from a randomized, controlled experiment at a US university, we study the relationships among SET, the gender of students, and the actual and perceived genders of instructors. The perceived gender of the instructor has a large, statistically significant effect (in the randomized experiment) or association (in the natural experiment) on SET: instructors perceived to be female receive lower SET scores. There are also statistically significant associations between SET and expected grades (higher expected grades are associated with higher SET scores). However, SET are not significantly associated with an objective measure of teaching effectiveness, student performance on anonymously graded standardized final exams. This suggests that SET do not primarily measure instructors' effects on student learning. Rather, they primarily measure measure student biases and expectations. There is strong statistical evidence that at the aggregate level, reliance

on SET for personnel decisions puts female instructors at a disadvantage. However, the extent and even the sign of gender biases vary by student and by subject, making it difficult or impossible for universities to adjust for such biases.

1 Background

Student evaluations of teaching (SET) are used widely as a measure of teaching quality, used in decisions about hiring, promoting, and firing instructors, especially non-tenured higher-education faculty. Universities generally treat SET as a measure of teaching effectiveness, rather than, e.g., a measure of student satisfaction. Because ascertaining teaching effectiveness is difficult—for students, faculty, and administrators alike—attempts to measure teaching effectiveness by surveying student opinion may suffer from conscious or unconscious biases. In this article, we adopt the definition by Centra and Gaubatz [2000, p.17], according to whom biases in SET correspond to a situation in which "a teacher or course characteristic affects teacher evaluations, either positively or negatively, but is unrelated to criteria of good teaching, such as increasedstudent learning."

Past research suggests that biases pervade SET. For instance, randomized experiments show that students confuse grades (or grade expectations) with long-term value [Carrell and West, 2010, Braga et al., 2014]. These experiments show that SET scores are not associated with student performance in follow-on courses, suggesting that SET do not measure actual teaching effectiveness nor effective learning. Instead, students may—on the whole—"reward" instructors who give higher grades for less effort. Instructors who are easy graders or who teach to the test may therefore obtain high SET scores, despite the fact that they encourage shallow learning.

Gender biases also affect how students rate instructors. Recent work by Boring [2015a] using data from a natural experiment at a French university suggests that gender biases and stereotypes influence SET. Male first-year undergraduate students tend to give more *excellent* scores to male instructors, even though an objective measure of student performance indicates that there is no difference between the

learning of students of male and of female instructors. Recent work by MacNell et al. [2014] involving a randomized, controlled experiment shows that, on average, students rate the very same instructor lower on every aspect of teaching, including "objective" measures such as timeliness, when they think the instructor is female than when they think the instructor is male.

Here, we use the datasets of Boring [2015a] and MacNell et al. [2014] to investigate hypotheses relating to whether the SET scores, including the overall satisfaction score, primarily measure teaching effectiveness or student biases. The two main sources of bias we study are students' grade expectations and the gender of the instructor. We also investigate systematic differences by discipline using data from Boring [2015a].

We use nonparametric permutation tests, which allow us to avoid counterfactual assumptions about generative models for the data, which regression-based methods (including ordinary linear regression, mixed effects models, logistic regression, etc.) and parametric methods such as t-tests and ANOVA, require. The null hypotheses for our tests are simply that some characteristic—e.g., instructor gender—amounts to an arbitrary label and might as well have been assigned at random.

Our analysis is conducted at the level of courses, matching how institutions use SET: typically, student responses in a given course are averaged, and those averages are compared across instances of the course, across courses in a department, across instructors, and across departments. Statistical problems with this reduction to and reliance upon averages are discussed by **?**

Associations between objective measures of teaching effectiveness and SET are weak and not statistically significant. Gender biases are stronger determinants of SET than teaching effectiveness is. Instructors perceived to be male receive significantly higher SET scores on average: in the French data, male students tend to

rate male instructors higher than they rate female instructors, with no difference in ratings by female students; in contrast, in the US data, *female* students tend to rate (perceived) male instructors higher than they rate female instructors, with no difference in ratings by male students. The French data also show that biases vary by course topic. The French data also that students conflate grade expectations with teaching effectiveness. We therefore conclude that SET primarily do not measure teaching effectiveness; that they have strong biases unrelated to actual teaching effectiveness; that the biases are not uniform; and that it is impossible to adjust for these biases. The strong gender biases and the weak association between SET and student performance imply that SET should not be relied upon as a measure of teaching effectiveness. In particular, reliance on SET for personnel decisions has disparate impact.

2 Methods

Throughout the paper, we use permutation tests to compare the SET of male and female instructors. Permutation tests suppose that we have a fixed number N of individuals randomized into two groups, a treatment group of size N_t and a control group of size $N_c = N - N_t$. For each of the N individuals, we measure a real-valued, quantitative response. Each individual i has two *potential* outcomes: their response t_i if they were assigned to the treatment group and their response c_i if they were assigned to the control group. We only observe one of these values for each individual, according to their treatment assignment. In our analyses, the "treatment" assigned at random to each student is instructor gender.

The *sharp* null hypothesis is that for all individuals, treatment has no effect:

$t_i = c_i$ for all $i = 1, \ldots, N$. This is a strong hypothesis in the sense that if it is true, we know both potential outcomes for all individuals. Namely, they would have the same response whether they were assigned to treatment or control. We assume that there is no interference between individuals: the treatment assignment of individual i does not affect the potential outcomes for individual j, for all pairs i and j.

Under this hypothesis, we can find the null distribution of any statistic computed from the responses of the two groups. The differences between statistics computed from the responses of the two groups are due to which individuals are randomly assigned to treatment and control. To generate this null distribution, we must generate all randomizations that were possible under the experimental design. In Boring [2015a], students sign up for triplets of classes without knowing the genders of their instructors, making it essentially random. In MacNell et al. [2014], students are assigned at random to one of four sections, making both their instructor's actual and reported gender a random treatment. Section 3.2 describes the randomization tests and choices of test statistics for the data from Boring [2015a]. Section 4.1 describes these for the data from MacNell et al. [2014].

The p-value of the test is the probability of observing a test statistic as extreme or more extreme under the null hypothesis. The p-value is obtained by comparing the observed test statistic to the null distribution. In practice, it is often too computationally intensive to find the null distribution exactly by enumerating all possible treatment assignments and computing the corresponding test statistic. Instead, we estimate the p-value by simulation. Let T^{obs} be the observed value of the test statistic. Then, assign treatment to N_t individuals at random by permuting the responses of all N individuals and taking the first N_t to be the treatment group. Note that

this is equivalent to generating a random treatment assignment vector of length N containing N_t ones and N_c zeros. For the complete randomization in Boring [2015a], each possible permutation has probability $1/\binom{N}{N_t}$. Calculate the test statistic for the permuted data, say T_1^*. Repeat this process of permuting the responses and calculating the test statistic $B-1$ more times. We then have an approximate null distribution of B test statistics $T_1^*, \ldots, T_B^*$ generated under the null hypothesis. For a one-sided test, where large values of T are "extreme", the estimated p-value is the proportion of these test statistics that are as large or larger than the observed test statistic:

$$\hat{p} = \frac{\{\#T_i^* \geq T^{obs}\}}{B}$$

For a two-sided test, if T is centered at 0 under the null, the estimated p-value is

$$\hat{p} = \frac{\{\#|T_i^*| \geq |T^{obs}|\}}{B}$$

In our implementations, we use $B = 10,000$ repetitions.

A benefit of the permutation approach to testing is that it is fully nonparametric. We make no distributional assumptions about the data or outcomes. Instead, the sharp null hypothesis fully specifies the distribution of any test statistic. The correctness of the p-value of a permutation test does not rely on any model specification, estimated parameters, or asymptotics. Another benefit of permutation tests is that we are free to choose any test statistic for our test; we can find the exact distribution of any test statistic, no matter how unusual, by simulation. A suitable test statistic should provide power to reject the null hypothesis in favor of some alternative hypothesis of interest. In this paper, most of our alternative hypotheses are that there

is a difference in the potential outcomes for some individuals, for various outcomes (e.g. SET and student grades). We use the difference in means between the two groups and the Spearman correlation as test statistics. These statistics are well-powered to detect a difference between two groups when their averages are different. Both of these test statistics have asymptotic distributions which only approximate the null distribution in finite samples. Our permutation approach allows us to find the null distributions of these statistics without relying on asymptotic results.

3 Tests of Boring [2015a]

3.1 The data

We first use a remarkable census of SET by first-year students at a French university, collected between 2008 and 2013, comprising 23,001 SET by 4,423 students (57% women) of 1,177 sections, taught by 379 instructors (34% women). These data are discussed in detail by Boring [2015a]. The key aspects of the data are as follows:

- All first year students take the same six mandatory courses, in: history, macroeconomics, microeconomics, political institutions, political science and sociology. Each course has one main professor, who delivers the lectures (to groups of approximately 900 students). All main professors are men. Courses have many sections of 10–24 students. Those sections are taught by different instructors. The instructors have considerable pedagogical freedom.

- Students enroll in "triads" of sections of these courses. The enrollment process does not allow students the freedom to select individual instructors. The assignment of students to sections is "as if" at random.

- Section instructors provide interim grades during the semester. Students know what their interim grades are when they complete their SET, so interim grades are a good measure of grade expectations.

- Final exams are created by the main professor, and all students in a given course take the same final. Final exams are graded anonymously in all disciplines except political institutions (which we omit from analyses involving final exam scores). This feature makes performance on the final exam a reasonable measure of the value the section instructor adds: students of more effective instructors should do better on the final exam, on average.

- SET are mandatory: the response rates are nearly perfect.

SET include closed-ended and open-ended questions, but the question that attracts the most attention is the overall score, which is considered to be a summary of the scores on the other questions.

Our SET database includes students' individual evaluations of instructors in the sections for microeconomics, history, political institutions, and macroeconomics for the five academic years 2008–2013, and for sociology and political science courses for the three academic years 2010–2013 (these two courses were introduced in 2010). The SET scores are anonymous to the instructors, who only have access to them once all grades have been officially recorded on student transcripts, several weeks after final exams. Instructors and academic coordinators then have access to the SET. When scores are low, the academic coordinator discusses the SET with the instructor and decides to maintain the instructor for the next semester or not.

While 34% of the 1,194 instruction sections are taught by women (Table 1), there is some variation by discipline. The political institutions sections are the ones more often taught by men (only 20% are taught by women). The sociology sections are

Table 1: Descriptive statistics of instruction sections

	N. courses	N. instructors	% Female instructors
Overall	**1,194**	**379**	**33.8%**
History	230	72	30.6%
Political institutions	229	65	20.0%
Microeconomics	230	96	38.5%
Macroeconomics	230	93	34.4%
Political science	137	49	32.7%
Sociology	138	56	46.4%

Note: the data for one political institutions section were excluded as this section had an experimental online format. The political science and sociology courses were originally not included in the triad system, and students were randomly assigned by the administration to different sections.

almost equally divided between male and female instructors (46.4% are taught by women). The microeconomics and macroeconomics sections have a larger number of different instructors, because of a higher turnover of instructors.

3.2 Methods

In this section, we carry out two sorts of analyses: tests for outcomes pertaining to instructors and tests pertaining to student behavior. In all our tests, we aggregate variables within instruction sections. Since students sign up for courses in triads without knowledge of who their instructors will be, random assignment of instructors to courses seems like a fair proxy. In these tests, we view each instruction section as an experimental unit, and the treatment each unit receives is their instructor assignment. With this randomization, an instructor is equally likely to be assigned to any course and its associated average SET and grades. Under the sharp null hypothesis, these variables do not change according to instructor, but are a property of the students in a course. We do not randomize individual students, as this would ignore possible correlations within courses. We use the Spearman correlation as the test

statistic.

We first examine the outcomes for each instructor. Course-level averages are relevant metrics that universities use to make hiring and tenure decisions. These include the correlation between average SET and average final grade (Table 2), average SET and average interim grade (Table 7), average SET and instructor gender (Table 3), and average final grade and instructor gender (Table 4). We perform these tests for all instructors in the data, assuming that they are exchangeable between departments, and also by department, allowing for the possibility that relationships between the variables of interest may vary by discipline.

Then, we examine how students of different genders rate their instructors and perform in classes. Stratifying by gender accounts for intrinsic differences between male and female students. It also allows us to identify interactions between student and instructor gender: if relationships between instructor gender and another outcome vary according to student gender, then there is an interaction. Here, we continue to aggregate student outcomes within courses, but now we take averages separately for the male and female students. These tests include the correlation between SET and gender concordance (Table 5) and the correlation between final exam grade and gender concordance (Table 6).

3.3 Analysis of SET and grades

Teaching effectiveness is multidimensional (e.g. Marsh and Roche [1997]) and is therefore difficult to measure. But effective teaching should generate student learning, suggesting that effective instructors should lead their students to understand

and learn more course material. Effective instructors should therefore cause their students to obtain higher grades on the final exams, on average.

We first test whether SET scores are correlated with higher grades on the final exam, on average by instruction section (Table 2). The results suggest that SET scores do not always measure actual teaching effectiveness. Overall, final exam grades are not statistically correlated with SET scores (one-sided p-value 0.70). The only two courses for which they are correlated are microeconomics and macroeconomics (p-values of 0.03 and 0.04). SET scores are uncorrelated with student achievement in the three other courses that are graded anonymously, i.e. history (p-value 0.31), political science (p-value 0.53) and sociology (p-value 0.27).

Table 2: Correlation between average SET scores and final exam grades, by instruction section

	ρ	p-value
Overall	-0.02	0.70
History	0.03	0.31
Macroeconomics	0.12	0.04
Microeconomics	0.13	0.03
Political science	-0.01	0.53
Sociology	0.05	0.27

Note: one-sided p-values are reported, since we expect that higher SET scores are likely to be correlated with higher final exam grades.

3.4 The correlation between SET scores and gender

Although mostly uncorrelated with students' performance on the final exam, SET appear to be much better predictors of gender. Overall, average SET scores and instructor gender are correlated, with male instructors obtaining significantly higher SET scores overall (p-value 0.00). There are, however, strong variations by course type (Table 3). Male instructors of history, macroeconomics and political

institutions courses receive (weakly) significantly higher overall satisfaction scores (p-values of 0.07, 0.08 and 0.10 respectively). The relationship is also positive between SET scores and instructor gender for microeconomics, political science and sociology courses, although not significant (p-values of 0.58, 0.43 and 0.26 respectively).

Table 3: Analyzing the correlation between average SET score and instructor gender, by course

	ρ	p-value
Overall	0.10	0.00
History	0.12	0.07
Political institutions	0.11	0.10
Macroeconomics	0.11	0.08
Microeconomics	0.04	0.58
Political sciences	0.07	0.43
Sociology	0.10	0.26

Note: two-sided p-values are reported.

Do men receive higher SET scores overall because they are better instructors? If men were indeed better instructors, then their students should perform better on final exams, on average, in comparable courses. This is not what we find (Table 4). Indeed, the correlation between student performance and instructor gender is negative, although statistically insignificant (p-value 0.51 overall), suggesting that male instructors are not more effective than female instructors, and perhaps are less effective.

So why do male instructors receive higher SET scores? SET scores and instructor gender are correlated, because male students tend to give higher SET scores to male instructors (Table 5). Our permutation tests confirm the results found by Boring [2015a]. Gender concordance is a statistically strong predictor of SET scores for men (p-value 0.00 overall). Male students give higher SET scores to male instructors in all fields. The correlations are statistically significant at level 0.1 in history (p-

Table 4: Correlation between final exam average and instructor gender, by course

	ρ	p-value
Overall	-0.02	0.51
History	-0.06	0.39
Macroeconomics	0.00	0.97
Microeconomics	-0.03	0.63
Political sciences	0.02	0.79
Sociology	-0.00	0.97

Note: two-sided p-values are reported.

value 0.00), macroeconomics (p-value 0.04), political science (p-value 0.06), political institutions (p-value 0.07) and microeconomics (p-value 0.10). The correlation is positive but not statistically significant in sociology (p-value 0.15).

Although gender concordance is correlated with overall satisfaction scores for male students, SET scores of female students are not statistically correlated with instructor gender (p-value 0.49 overall). The correlation is negative in some fields (history, political institutions, macroeconomics and sociology) and positive in others (microeconomics and political science), but always statistically insignificant (p-values between 0.19 and 0.97).

Table 5: Correlation between SET scores and gender concordance

	Male student		Female student	
	ρ	p-value	ρ	p-value
Overall	0.15	0.00	0.02	0.49
History	0.18	0.00	-0.04	0.54
Political institutions	0.12	0.07	-0.09	0.19
Macroeconomics	0.14	0.04	-0.08	0.21
Microeconomics	0.11	0.10	0.03	0.67
Political sciences	0.16	0.06	0.00	0.97
Sociology	0.12	0.15	-0.05	0.53

Note: two-sided p-values are reported.

Do male instructors receive higher SET scores from male students because their

teaching styles match male students' learning styles? If that were the case, then male students who had male instructors should perform better on the final exam. However, this is not what we find (Table 6). If anything, male students who had male instructors appear to perform worse overall on the final exam (the correlation is negative but statistically insignificant, with a p-value 0.76). In history, the negative correlation (-0.11) is weakly statistically significant (p-value 0.10). In history, male students therefore give significantly higher SET scores, despite the fact that they appear to learn more from female instructors. These results further suggest that students are not measuring actual teaching effectiveness when they complete their SET.

Table 6: Student performance and gender concordance

	Male student		Female student	
	ρ	p-value	ρ	p-value
Overall	-0.01	0.76	0.01	0.65
History	-0.11	0.10	0.01	0.86
Macroeconomics	0.02	0.76	-0.00	0.97
Microeconomics	-0.04	0.60	0.00	0.94
Political sciences	0.10	0.25	0.03	0.76
Sociology	0.02	0.85	-0.01	0.94

Note: two-sided p-values are reported.

3.5 The correlation between SET scores and grade expectations

Not only are SET scores correlated with gender, but they are also positively and significantly correlated with expected grades (Table 7). Political institutions is the only course for which the correlation between expected grades and SET scores is not significant (p-value 0.19). The p-values in all other courses are close to 0.

The correlation coefficients are especially high in history (0.32) and sociology (0.27). They are also high in macroeconomics (0.22), microeconomics (0.19) and political sciences (0.16).

Table 7: Analyzing the correlation between average evaluation score and interim grades, by course number

	ρ	p-value
Overall	0.10	0.00
History	0.32	0.00
Political institutions	0.06	0.19
Macroeconomics	0.22	0.00
Microeconomics	0.19	0.00
Political sciences	0.16	0.03
Sociology	0.27	0.00

Note: one-sided p-values are reported.

To summarize our results, the fact that SET scores are largely uncorrelated with student achievement measured by students' final exam grades suggests that (male) students may be expressing a gender bias in favor of men when rating instructors. Furthermore, students appear to reward instructors who give them higher interim grades. We conclude that gender and expected grades create biases in SET scores, which are unrelated to effective teaching.

4 Tests of MacNell et al. [2014]

While our analysis of the data in the previous section suggests that SET scores are largely unrelated to teaching effectiveness, the natural experimental setting of the French university data does not enable us to control for potential differences in teaching styles of men and women. We know of two experiments which were able to control for teaching styles: Arbuckle and Williams [2003] and MacNell et al. [2014].

These two experiments tend to confirm that students express gender biases in SET scores, rather than reward a teaching style that matches their learning style. In both experiments, students tend to give higher SET scores when they think that the course is being taught by a man, regardless of whether the course is actually taught by a man or a woman. Hence, differences in teaching or learning styles do not seem to explain the differences in men and women's SET scores.

In the Arbuckle and Williams [2003] experiment, a group of 352 students watched "slides of an age- and gender-neutral stick figure and listened to a neutral voice presenting a lecture and then evaluated it on teacher evaluation forms that indicated 1 of 4 different age and gender conditions (male, female, "old," and "young")" [Arbuckle and Williams, 2003, p.507]. The goal of the experiment was to measure whether "students' perceptions of a professor's age and gender influence their perceptions of the professor's warmth and enthusiasm". Differences in evaluations could thus only be caused by students' subjective age and gender-biased judgments in evaluating the lecturer's competence. The researchers find that students rated the young male instructors higher than the other three combinations, especially on "enthusiasm", "showed interest in subject" and "using a meaningful voice tone".

The results of Arbuckle and Williams [2003] are reinforced by those of MacNell et al. [2014] who use a different set-up to control for differences in teaching styles. In their experiment, MacNell et al. [2014] used SET data collected from an online course in which 43 students were randomly assigned to four discussion groups for a course, each taught by one of two assistant instructors (one man and one woman). The two instructors each taught one course under their real identity, while they taught the other course under the other instructor's identity. In this set-up, one group of twelve students thought they had a female instructor when the instructor was actually male, and twelve other students thought they had the male instructor when the instructor

was actually female. The two instructors worked together with the main professor of the course, to make sure that they gave similar types of feedback to students, graded papers in exactly the same time frame, etc., so as to limit differences in teaching styles or grading to a strict minimum.

With this framework, potential gender biases can be tested by controlling for teaching styles. Biases in student ratings can be found by comparing how students rate their instructors as a function of the *actual* versus *perceived* gender of the instructor. MacNell et al. [2014] find that "the male identity received significantly higher scores on professionalism, promptness, fairness, respectfulness, enthusiasm, giving praise, and the student ratings index... [...] Students in the two groups that perceived their assistant instructor to be male rated their instructor significantly higher than did the students in the two groups that perceived their assistant instructor to be female, regardless of the actual gender of the assistant instructor."

4.1 Methods

In this section, we once again use permutation tests, this time to analyze the data provided by MacNell et al. [2014]. The use of permutation tests is especially appropriate in this case, given the small sample of twenty male and twenty-three female students.

Each student's potential responses are represented by numbers on a ticket:

- the rating that the student would assign to instructor 1 if instructor 1 is identified as male
- the rating that the student would assign to instructor 1 if instructor 1 is identified as female

- the rating that the student would assign to instructor 2 if instructor 2 is identified as male
- the rating that the student would assign to instructor 2 if instructor 2 is identified as female

The null hypothesis is that the first two numbers are equal and the second two numbers are equal, but the first two numbers might be different from the second two numbers. This corresponds to the hypothesis that students assigned to a given instructor would rate him or her the same, whether that instructor seemed to be male or female. For all students assigned to instructor 1, we know both of the first two numbers if the null hypothesis is true, but we know neither of the second two numbers. Similarly, if the null hypothesis is true, we know both of the second two numbers for all students assigned to instructor 2, but we know neither of the first two numbers for those students.

Because of how the randomization was performed, all allocations of students to class sections that preserve the number of students in each section are equally likely. In particular, all allocations that keep the same students assigned to each actual instructor the same are equally likely.

To test the difference in SET between male-identified and female-identified instructors, we use the difference in means between the two groups as our test statistics (Table 8). We look at the instructors' overall rating as well as their rating in each category. To approximate the null distributions, we permute students who were assigned to instructor 1 and instructor 2 separately, then allocate students within those two groups to the male-identified and female-identified sections. This stratified

permutation method corresponds to testing the null described above.

Additionally, we test the correlation between ratings and the concordance of student and reported instructor gender (Table 9), correlation between ratings and concordance of student and actual instructor gender (Table 10), and correlation between student grades and actual instructor gender(Table 11). For these tests, we use the Spearman correlation. The tests involving reported instructor gender use the stratified permutations described above, where as the tests of actual instructor gender simply permute all students irrespective of their instructor's reported gender.

4.2 Analysis of SET and gender

We first analyze the correlation between student ratings and the reported instructor gender (Table 8). A positive result signifies that the perceived male instructor received higher evaluations. We find a weak positive correlation between the perceived gender and overall satisfaction (p-value 0.10). The statistical significance is stronger for several of the criteria which students rated: fairness (p-value 0.00), giving praise (p-value 0.01), caring and promptness (both criteria have p-values of 0.04), enthusiasm (p-value 0.05), communication (p-value 0.06), professionalism and respect (both criteria have p-values of 0.07), and being consistent and helpful (both criteria have p-values of 0.09). The criteria that were not statistically significant were feedback, responsiveness, being knowledgeable and clear. Our permutation tests confirm and extend the results found by MacNell et al. [2014].

We then analyze in more detail whether male or female students rated the instructors differently according to perceived gender. While in the previous section we found that male students rated male instructors higher, we find in the Mac-

Table 8: Analyzing the difference in mean ratings and reported instructor gender (male minus female)

	difference in means	p-value
Overall	0.47	0.15
Professional	0.61	0.06
Respectful	0.61	0.06
Caring	0.52	0.12
Enthusiastic	0.57	0.08
Communicate	0.57	0.09
Helpful	0.46	0.19
Feedback	0.47	0.19
Prompt	0.80	0.02
Consistent	0.46	0.24
Fair	0.76	0.01
Responsive	0.22	0.54
Praise	0.67	0.02
Knowledge	0.35	0.29
Clear	0.41	0.34

Note: two-sided p-values are reported.

Nell et al. [2014] experiment that the perceived male instructor received significantly higher evaluation scores because female students rated the perceived female instructor significantly lower (Table 9). Male students rated the perceived male instructor significantly (though weakly) higher on only one criteria: being fair (p-value 0.08). Female students, however, rated the perceived female instructor lower in terms of overall satisfaction (p-values of 0.08), along most teaching dimensions: giving praise (p-value 0.01), enthusiasm (p-value 0.03), caring and fairness (p-values of 0.04), being respectful and communication (p-values of 0.08), professionalism (p-value 0.09) and feedback (p-value 0.10). Although the results show a negative correlation between being a (perceived) female instructor and ratings on being helpful, promptness, consistency, responsiveness, knowledge and clarity, the results are not statistically significant.

When we analyze how students rated the instructors according to their *actual*

Table 9: Analyzing the correlation between ratings and reported instructor gender, by gender concordance

	Both male		Both female	
	ρ	p-value	ρ	p-value
Overall	0.09	0.81	-0.36	0.11
Professional	0.22	0.52	-0.36	0.10
Respectful	0.22	0.34	-0.36	0.10
Caring	0.02	1.00	-0.46	0.05
Enthusiastic	0.09	0.82	-0.44	0.05
Communicate	0.12	0.68	-0.39	0.10
Helpful	0.21	0.41	-0.24	0.35
Feedback	0.04	0.90	-0.37	0.10
Prompt	0.38	0.15	-0.37	0.13
Consistent	0.07	0.85	-0.34	0.18
Fair	0.41	0.09	-0.43	0.04
Responsive	0.18	0.53	-0.03	0.99
Praise	0.29	0.27	-0.47	0.01
Knowledge	0.08	0.78	-0.29	0.21
Clear	0.06	0.76	-0.25	0.29

Note: two-sided p-values are reported.

Table 10: Analyzing the correlation between ratings and actual instructor gender, by gender concordance

	Both male		Both female	
	ρ	p-value	ρ	p-value
Overall	-0.07	0.72	0.13	0.56
Professional	0.08	0.74	0.04	0.95
Respectful	0.08	0.84	0.04	0.94
Caring	-0.11	0.59	0.03	0.98
Enthusiastic	-0.07	0.82	0.20	0.40
Communicate	-0.01	0.84	0.08	0.67
Helpful	0.01	0.96	-0.12	0.70
Feedback	-0.12	0.69	0.17	0.50
Prompt	-0.05	0.89	0.14	0.53
Consistent	0.05	0.85	0.17	0.49
Fair	-0.03	0.88	0.28	0.23
Responsive	-0.06	0.84	0.35	0.12
Praise	0.01	1.00	0.34	0.13
Knowledge	0.11	0.70	0.24	0.36
Clear	-0.12	0.65	0.35	0.12

Note: two-sided p-values are reported.

gender, we find no significant difference in evaluations (Table 10). We do find, however, that the students of the actual male instructor performed better in the course and obtained significantly higher grades (Table 11). There is no statistical difference between student performance and the perceived gender of the instructor.

Table 11: Correlation between grade and instructor gender

	t-stat	p-value
Perceived	0.21	0.83
Actual	2.65	0.01

Note: two-sided p-values are reported.

These results suggest that students did not rate the two instructors as a function of their actual teaching effectiveness (which in this experiment may be confounded with gender, i.e. the actual male instructor being a better instructor than the female instructor). Instead, students appear to have rated instructors largely as a function of the perceived gender of the instructor. Our analysis suggests that the female students were biased against the *perceived* female instructor, but were unable to tell the difference between the *actual* male and female instructor.

5 Code

Github repo. https://github.com/kellieotto/SET-and-Gender-Bias

6 Conclusions

Teaching effectiveness is a vague notion that even researchers of higher education have a hard time defining. There is a consensus that teaching effectiveness is multidimensional (e.g. [Marsh and Roche, 1997]), and that universities must find

incentives to encourage better teaching. The notion of teaching effectiveness implies that instructors have some control over the impact their teaching skills have on student-related outcomes. Measures of teaching effectiveness should therefore only reflect variables that are under the control of instructors.

In our analysis, we used a robust statistical test to develop the results by Boring [2015a] and MacNell et al. [2014], which suggest that gender biases prevent SET from objectively measuring teaching effectiveness. Our results confirm that an instructor's perceived gender may be more important to students in the way they rate instructors, than student-related outcomes such as an instructor's ability to help student learning. Instructors appear to be rated to a larger extent on a variable that is out of their control (their gender), rather than their ability to positively impact student learning. We further find that the extent and direction of gender biases appear to depend on context. While the French university setting highlights a positive male student bias for male instructors, the experimental US setting suggests a negative female student bias against female instructors.

Instead of measuring teaching effectiveness, SET appear to be a measure of student satisfaction regarding a course [**?**]. Students may be satisfied or dissatisfied with courses for reasons outside of the control of instructors. Gender may be one of these reasons, due to a given cultural context for example. We do find that the correlation between SET and performance isn't zero: it can be positive, albeit context dependent and not always statistically significant. While student satisfaction can be considered to be one dimension of teaching effectiveness, the larger point of our analysis is that SET are better measures of student grade expectations and of instructor gender than they are of teaching effectiveness.

Gender and expected grades are not the only variables unrelated to teaching effectiveness that other studies have shown to be predictors of SET scores. Given

the many variables that are likely to bias SET scores and whose weight in SET are likely to change from one learning environment to another, it would be impossible to control for all these variables to make SET a valid measure of teaching effectiveness. Furthermore, the direction of biases appear to be context dependent.

Among the instructor characteristics alongside gender, race has also been shown to be correlated with SET scores. In studies conducted in the US, instructors of color appear to suffer from student biases similar to those that female instructors suffer from in our analysis. Minority instructors tend to receive significantly lower SET scores compared to white (male) instructors (e.g. Merritt [2008]).[1] Other instructor-related characteristics likely to be unrelated to teaching effectiveness have been shown to be predictors of SET scores, such as age [Arbuckle and Williams, 2003], charisma [Shevlin et al., 2000] and physical attractiveness (e.g. Riniolo et al. [2006] and Hamermesh and Parker [2005]).

Other factors still unrelated to factors that an instructor can control may be related to SET scores. Variables related to the teaching environment, class time, class size, mathematical content of the course, etc. may matter. For instance, Hill and Epps [2010] show that students' perceptions of classroom environment factors (such as seating characteristics or lighting) have an impact on student ratings of instructors. They find that differences in the physical characteristics of classrooms influence students' overall satisfaction with a course, and have an impact on student evaluations of criteria such as their perceptions of how organized their instructors are.

Hundreds of studies discuss and question the validity of SET as a measure of teaching effectiveness (e.g. for reviews Pounder [2007]). Some studies find results

[1]French law does not allow for the use of race-related variables in data sets. We were thus unable to test for potential racial biases in SET scores in the context of our French university.

that are similar to ours, with male students expressing biases in favor of male instructors (e.g. Basow and Silberg [1987]; Kaschak [1978]). Other studies find that the gender and SET is uncorrelated or that the relationship is weak (e.g. Bennett [1982]; Centra and Gaubatz [2000]; Elmore and LaPointe [1974]). While some studies tend to suggest that SET are not a valid measure of teaching effectiveness e.g. Galbraith et al. [2012] and Carrell and West [2010]), others argue that SET are valid and reliable measures of teaching effectiveness (e.g. Benton and Cashin [2012] and Centra [1977]). While there is no consensus among academics on the issue of validity, the fact that different studies show such a wide variety of results suggests that validity varies with contexts. This fact, in itself, shows that SET are not universally valid and should be used by universities with great caution.

In the US, SET have two primary purposes: to help instructors improve their teaching and to help the administration make personnel decisions, such as hiring or promoting instructors. We recommend discontinuing the second use of SET, given the strong student biases that influence SET, even on "objective" items. In fact, in France, the French Ministry of Higher Education and Research upheld in 2009 a 1997 decision of the French State Council that public universities can use SET only to help tenured instructors improve their pedagogy, and that the administration may not use SET in decisions that might affect tenured instructors' careers (cf. Boring [2015b]).

Our results suggest that the existence of gender biases in SET is context dependent. To test for the external validity of our results, we encourage the replication of our analysis in different settings. The results we find suggest that, in some contexts, female instructors may receive lower than average SET scores, despite being as effective instructors as men, only because of student biases in favor of male instructors. The use of SET therefore unfairly penalizes women, and can have large consequences

on their academic careers. Our results more generally emphasize that, at least in some contexts, instructors are being unfairly judged based on variables that are out of their control, potentially leading to negative consequences on their careers in academia. We encourage universities to study potential biases that may occur in their contexts, and to take appropriate measures so as to not penalize instructors for variables that are out of their control.

References

J. Arbuckle and B. D. Williams. Students ' Perceptions of Expressiveness : Age and Gender Effects on Teacher Evaluations. *Sex Roles*, 49(November):507–516, 2003.

S. A. Basow and N. T. Silberg. Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology*, 79(3): 308–314, 1987.

S. K. Bennett. Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, 74(2):170–179, 1982.

S. L. Benton and W. E. Cashin. Student ratings of teaching: A summary of research and literature. IDEA Paper 50, The IDEA Center, 2012.

A. Boring. Gender biases in student evaluations of teachers. Document de travail OFCE 13, OFCE, April 2015a.

A. Boring. Can students evaluate teaching quality objectively? Le blog de l'ofce, OFCE, 2015b. URL http://www.ofce.sciences-po.fr/blog/can-students-evaluate-teaching-quality-objectively/.

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

S. E. Carrell and J. E. West. Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *Journal of Political Economy*, 118(3):409–432, June 2010. ISSN 0022-3808. doi: 10.1086/653808. URL http://www.jstor.org/stable/10.1086/653808.

J. A. Centra. Student ratings of instruction and their relationship to student learning. *American educational research journal*, 14(1):17–24, 1977.

J. A. Centra and N. B. Gaubatz. Is There Gender Bias in Student Evaluations of Teaching? *Journal of Higher Education*, 71(1):17–33, 2000. URL http://www.jstor.org/stable/10.2307/2649280.

P. B. Elmore and K. A. LaPointe. Effects of teacher sex and student sex on the evaluation of college instructors. *Journal of Educational Psychology*, 66(3):386–389, 1974.

C. S. Galbraith, G. B. Merrill, and D. M. Kline. Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? a neural network and bayesian analyses. *Research in Higher Education*, 53(3):353–374, 2012.

D. S. Hamermesh and A. Parker. Beauty in the classroom: Instructors pulchritude and putative pedagogical productivity. *Economics of Education Review*, 24(4): 369–376, 2005.

M. C. Hill and K. K. Epps. The impact of physical classroom environment on stu-

dent satisfaction and student evaluation of teaching in the university environment. *Academy of Educational Leadership Journal*, 14(4):65–79, 2010.

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

L. MacNell, A. Driscoll, and A. N. Hunt. Whats in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, pages 1–13, 2014.

H. W. Marsh and L. A. Roche. Making Students' Evaluations of Teaching Effectiveness Effective. *American Psychologist*, 52(11):1187–1197, 1997.

D. J. Merritt. Bias, the brain, and student evaluations of teaching. *St. John's Law Review*, 81(1):235–288, 2008.

J. S. Pounder. Is student evaluation of teaching worthwhile?: An analytical framework for answering the question. *Quality Assurance in Education*, 15(2):178–191, 2007. ISSN 0968-4883. doi: 10.1108/09684880710748938. URL http://www.emeraldinsight.com/10.1108/09684880710748938.

T. C. Riniolo, K. C. Johnson, T. R. Sherman, and J. A. Misso. Hot or not: do professors perceived as physically attractive receive higher student evaluations? *The Journal of general psychology*, 133(1):19–35, Jan. 2006. ISSN 0022-1309. doi: 10.3200/GENP.133.1.19-35. URL http://www.ncbi.nlm.nih.gov/pubmed/16475667.

M. Shevlin, P. Banyard, M. Davies, and M. Griffiths. The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assessment & Evaluation in Higher Education*, 25(4):397–405, 2000.

P. B. Stark and R. Freishtat. An evaluation of course evaluations. *Science Open Research*, 2014. URL https://www.scienceopen.com/document/vid/42e6aae5-246b-4900-8015-dc99b467b6e4?10.