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

Anne Boring

Department of Economics

SciencesPo, Paris

Kellie Ottoboni and Philip B. Stark

Department of Statistics

University of California, Berkeley

Draft January 3, 2016

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

Gloria Steinem

Abstract

Student evaluations of teaching (SET) are widely used in academic personnel decisions as a measure of teaching effectiveness. We show:

- SET are biased against female instructors by an amount that is large and statistically significant
- the bias affects how students rate even putatively objective aspects of teaching, such as how promptly assignments are graded
- the bias varies by discipline and by student gender, among other things
- it is not possible to adjust for the bias, because it depends on so many factors
- SET are more sensitive to students' gender bias and grade expectations than they are to teaching effectiveness
- gender biases can be large enough to cause more effective instructors to get lower SET than less effective instructors

Relying on SET for personnel decisions disadvantages female instructors, and should be abandoned. These findings are based on nonparametric statistical tests applied to two datasets: 23,001 SET of 379 instructors by 4,423 students in six mandatory first-year courses in a five-year natural experiment at a French university, and 43 SET for four sections of an online course in a randomized, controlled, blind experiment at a US university.

1 Background

Student evaluations of teaching (SET) are used widely in decisions about hiring, promoting, and firing instructors. Measuring teaching effectiveness is difficult—for students, faculty, and administrators alike. Universities generally treat SET as if they primarily measure teaching effectiveness or teaching quality. While it may seem natural to think that students' answers to questions like "how effective was the instructor?" measure teaching effectiveness, it is not a foregone conclusion that they do. Indeed, the best evidence so far shows that they do not: they have *biases*[1] that are stronger than any connection they might have with effectiveness. Worse, in some circumstances the association between SET and an objective measure of teaching effectiveness is *negative*, as our results below reinforce.

Randomized experiments [Carrell and West, 2010, Braga et al., 2014] have shown that students confuse grades and grade expectations with the long-term value of a course and that SET are not associated with student performance in follow-on courses, a proxy for teaching effectiveness. On the whole, high SET seem to be a reward students give instructors who make them anticipate getting a good grade, for whatever reason; for extensive discussion, see Johnson [2003, Chapters 3–5].

Gender matters too. Boring [2015a] finds that SET are affected by gender biases and stereotypes. Male first-year undergraduate students give more *excellent* scores to male instructors, even though there is no difference between the academic performance of male students of male and of female instructors. Experimental work by MacNell et al. [2014] finds that when students think an instructor is female, students rate the instructor lower on every aspect of teaching, including putatively objective

[1] Centra and Gaubatz [2000, p.17] define bias to occur when "a teacher or course characteristic affects teacher evaluations, either positively or negatively, but is unrelated to criteria of good teaching, such as increased student learning."

measures such as the timeliness with which instructors return assignments.

Here, we apply nonparametric permutation tests to data from Boring [2015a] and MacNell et al. [2014] to investigate whether SET primarily measure teaching effectiveness or biases using a higher level of statistical rigor. The two main sources of bias we study are students' grade expectations and the gender of the instructor. We also investigate variations in bias by discipline and by student gender.

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

We work with course-level summaries to match how institutions use SET: typically, SET are averaged for each offering of a course, and those averages are compared across instances of the course, across courses in a department, across instructors, and across departments. Stark and Freishtat [2014] discuss statistical problems with this reduction to and reliance upon averages.

We find that the association between SET and an objective measure of teaching effectiveness, performance on the anonymously graded final, is weak and—for these data—generally not statistically significant. In contrast, the association between SET and (perceived) instructor gender is large and statistically significant: instructors whom (students believe) are male receive significantly higher average SET.

In the French data, *male* students tend to rate male instructors higher than they rate female instructors, with little difference in ratings by female students. In the US data, *female* students tend to rate (perceived) male instructors higher than they rate (perceived) female instructors, with little difference in ratings by male students.

The French data also show that gender biases vary by course topic, and that SET have a strong positive association with students' grade expectations.

We therefore conclude that SET primarily do not measure teaching effectiveness; that they are strongly and non-uniformly biased by factors including the genders of the instructor and student; that they disadvantage female instructors; and that it is impossible to adjust for these biases. SET should not be relied upon as a measure of teaching effectiveness. Relying on SET for personnel decisions has disparate impact by gender, in general.

2 Data

2.1 French Natural Experiment

These data, collected between 2008 and 2013, are a census of 23,001 SET from 4,423 first-year students at a French university students (57% women) in 1,177 sections, taught by 379 instructors (34% women). The data are not public, owing to French restrictions on human subjects data. Boring [2015a] describes the data in detail. Key features include:

- All first-year students take the same six mandatory courses: History, Macroeconomics, Microeconomics, Political Institutions, Political Science, and Sociology. Each course has one (male) professor who delivers the lectures to groups of approximately 900 students. Courses have sections of 10–24 students. Those sections are taught by a variety of instructors, male and female. The instructors have considerable pedagogical freedom.

- Students enroll in "triads" of sections of these courses (three courses per semester). The enrollment process does not allow students to select individual instructors.

The assignment of instructors to sets of students is as if at random, forming a *natural experiment*. It is reasonable to treat the assignment as if it is independent across courses.

- Section instructors assign interim grades during the semester. Interim grades are known to the students before the students submit SET. Interim grades are thus a proxy for students' grade expectations.

- Final exams are written by the course professor, not the section instructors. Students in all sections of a course in a given year take the same final. Final exams are graded anonymously, except in Political Institutions, which we therefore omit from analyses involving final exam scores. To the extent that the final exam measures appropriate learning outcomes, performance on the final is a measure of the effectiveness of an instructor: in a given course in a given year, students of more effective instructors should do better on the final exam, on average, than students of less effective instructors.

- SET are mandatory: response rates are nearly 100%.

SET include closed-ended and open-ended questions. The item that attracts the most attention, especially from the administration, is the *overall score*, which is treated as a summary of the other items. The SET data include students' individual evaluations of section instructors in microeconomics, history, political institutions, and macroeconomics for the five academic years 2008–2013, and for political science and sociology for the three academic years 2010–2013 (these two subjects were introduced in 2010). The SET are anonymous to the instructors, who have access to SET only after all grades have been officially recorded.

Table 1: Summary statistics of sections

course	# sections	# 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%

Data for a section of Political Institutions that had an experimental online format are omitted. Political Science and Sociology originally were not in the triad system; students were randomly assigned by the administration to different sections.

2.2 US Randomized Experiment

These data, described in detail by MacNell et al. [2014], are available at `http://n2t.net/ark:/b6078/d1mw2k`. Students in an online course were randomized into six sections of about a dozen students each, two taught by the primary professor, two taught by a female graduate teaching assistant (TA), and two taught by a male TA. In one of the two sections taught by each TA, the TA used her or his true name; in the other, she or he used the other TA's identity. Thus, in two sections, the students were led to believe they were being taught by a woman and in two they were led to believe they were being taught by a man. Students had no direct contact with TAs: the primary interactions were through online discussion boards. The TA credentials presented to the students were comparable; the TAs covered the same material; and assignments were returned at the same time in all sections (hence, objectively, the TAs returned assignments equally promptly in all four sections).

SET included an overall score and questions relating to professionalism, respectfulness, care, enthusiasm, communication, helpfulness, feedback, promptness, consistency, fairness, responsiveness, praise, knowledge, and clarity. Fourty-seven students in the four sections taught by TAs finished the class, of whom 43 submitted SET.

The SET data include the genders and birth years of the students;[2] the grade data do not. The SET data are not linked to the grade data.

3 Methods

Previous analyses of these data relied on parametric tests based on null hypotheses that do not match the experimental design. For example, the tests assumed that SET of male and female instructors are independent random samples from normally distributed populations with equal variances and possibly different means. As a result, the p-values reported in those studies are for unrealistic null hypotheses and might be misleading.

In contrast, we use permutation tests based on the as-if-random (French natural experiment) or truly random (US experiment) assignment of students to class sections, with no counterfactual assumption that the students, SET scores, grades, or any other variables comprise random samples from any populations, much less populations with normal distributions.

In most cases, our tests are *stratified*. For the US data, for instance, the randomization is stratified on the actual TA: students are randomized within the two sections taught by each TA, but students assigned to different TAs comprise different strata. The randomization is independent across strata. For the French data, the randomization is stratified on course and year: students in different courses or in different years comprise different strata, and the randomization is independent across strata. The null distributions of the test statistics[3] are induced by this random assignment, with no assumption about the distribution of SET or other variables, no parameter

[2] One birth year is obviously incorrect, but our analyses do not rely on the birth years.

[3] The test statistics are correlations of a response variable with experimental variables, or differences in the means of a response variable across experimental conditions, aggregated across strata.

estimates, and no model.

3.1 Illustration: French natural Experiment

The selection of course sections by students at the French university—and the implicit assignment of instructors to sets of students—is as if at random within sections of each course each year, independent across courses and across years. The university's triad system groups students in their classes across disciplines, building small cohorts for each semester. Hence, the randomization for our test keeps these groups of students intact. Stratifying on course topic and year keeps students who took the same final exam grouped in the randomization and honors the design of the natural experiment.

Teaching effectiveness is multidimensional [Marsh and Roche, 1997] and difficult to define, much less measure. But whatever it is, effective teaching should promote student learning: *ceteris paribus*, students of an effective instructor should have better learning outcomes than students of an ineffective instructor have. In the French university, in all courses other than Political Institutions,[4] students in every section of a course in a given year take the same anonymously graded final exam. To the extent that final exams are designed well, scores on these exams reflect relevant learning outcomes for the course. Hence, in each course each semester, students of more effective instructors should do better on the final, on average, than students of less effective instructors.

Consider testing the hypothesis that SET are unrelated to performance on the final exam against the alternative that, all else equal, students of instructors who get higher average SET get higher final exam scores, indicating that they learned more.

[4]The final exam in Political Institutions is oral and hence not graded anonymously.

For this hypothesis test, we omit Political Institutions because the final exam was not anonymous.

The test statistic is the average over courses and years of the Pearson correlation between mean SET and mean final exam score among sections of each course each year. If SET do measure instructors' contributions to learning, we would expect this average correlation to be positive: sections with above-average mean SET in each discipline each year would tend to be sections with above-average mean final exam scores. How surprising is the observed average correlation, if there is no overall connection between mean SET and mean final exam for sections of a course?

There are 950 "individuals," course sections of subjects other than Political Institutions. Each of the 950 course sections has an average SET and an average final exam score. These fall in $3 \times 5 + 2 \times 3 = 21$ year-by-course strata. Under the randomization, within each stratum, instructors are assigned sections independently across years and courses, with the number of sections of each course that each instructor teaches each year held fixed. For instance, if in 2008 there were N sections of History taught by K instructors in all, with instructor k teaching N_k sections, then in the randomization, all

$$\binom{N}{N_1 \cdots N_K} \tag{1}$$

ways of assigning N_k of the N 2008 History sections to instructor k, for $k = 1, \ldots, K$, would be equally likely. The same would hold for sections of other courses and other years. Each combination of assignments across courses and years is equally likely: the assignments are independent across strata.

Under the null hypothesis that SET have no relationship to final exam scores, average final exam scores for sections in each course each year are *exchangeable* given the average SET for the sections. Imagine "shuffling" (i.e., permuting) the average

final exam scores across sections of each course each year, independently for different courses and different years. For each permutation, compute the Pearson correlation between average SET for each section and average final exam score for each section, for each course, for each year. Average the resulting 21 Pearson correlations. The probability distribution of that average is the null distribution of the test statistic. The p-value is the upper tail probability beyond the observed value of the test statistic, for that null distribution.

The hypothetical randomization holds triads fixed, to allow for cohort effects and to match the natural experiment. Hence, the test is conditional on which students happen to sign up for which triad. However, if we test at level no greater than α conditionally on the grouping of students into triads, the unconditional level of the resulting test across all possible groupings is no greater than α:

$$\begin{aligned}
\Pr\{\text{ Type I error }\} &= \sum_{\text{all possible sets of triads}} \Pr\{\text{ Type I error } | \text{ triads }\} \Pr\{\text{ triads }\} \\
&\leq \sum_{\text{all possible sets of triads}} \alpha \Pr\{\text{triads }\} \\
&= \alpha \sum_{\text{all possible sets of triads}} \Pr\{\text{triads }\} \\
&= \alpha. \qquad (2)
\end{aligned}$$

It is not practical to enumerate all possible permutations of sections within courses and years, so we estimate the p-value by performing 10^5 random permutations within each stratum, finding the value of the test statistic for each overall assignment, and comparing the observed value of the test statistic to the empirical distribution of those 10^5 random values. The probability distribution of the number of random permutations assignments for which the test statistic is greater than or

equal to its observed value is Binomial, with n equal to the number of overall random permutations and p equal to the true p-value. Hence, the standard error of the estimated p-values is hence no larger than $(1/2)/\sqrt{10^5} \approx 0.0016$. Code for all our analyses is at https://github.com/kellieotto/SET-and-Gender-Bias. Results for the French data are below in section 4.

3.2 Illustration: US Experiment

To test whether perceived instructor gender affects SET in the US experiment, we use the Neyman "potential outcomes" framework [Neyman et al., 1990]. A fixed number N of individuals—e.g., students or classes—are assigned randomly (or as if at random by Nature) into $k \geq 2$ groups of sizes $N_1, \ldots, N_k$. Each group receives a different treatment. "Treatment" is notional. For instance, the treatment might be the gender of the class instructor.

For each individual i, we observe a numerical response R_i. If individual i is assigned to treatment j, then $R_i = r_{ij}$. The numbers $\{r_{ij}\}$ are considered to have been fixed before the experiment. (They are not assumed to be a random sample from any population; they are not assumed to be realizations of any underlying random variables.) Implicit in this notation is the *non-interference* assumption that each individual's response depends only on the treatment that individual receives, and not on which treatments other individuals receive.

We observe only one potential outcome for individual i, depending on which treatment she or he receives. In this model, the responses $\{R_i\}_{i=1}^N$ are random, but only because individuals are assigned to treatments at random, and the assignment determines which of the fixed values $\{r_{ij}\}$ are observed.

In the experiment conducted by MacNell et al. [2014], N students were assigned

at random to six sections of an online course, of which four were taught by TAs. Our analysis focuses on the four sections taught by TAs. We condition on the assignment of students to the two sections taught by the professor. Each remaining student i could be assigned to any of $k = 4$ treatment conditions: either of two TAs, each identified as either male or female. The assignment of students to sections was random: each of the

$$\binom{N}{N_1 N_2 N_3 N_4} = \frac{N!}{N_1! N_2! N_3! N_4!} \tag{3}$$

possible assignments of N_1 students to TA 1 identified as male, N_2 student to TA 1 identified as female, etc., was equally likely.

Let r_{i1} and r_{i2} be the ratings student i would give TA 1 when TA 1 is identified as male and as female, respectively; and let r_{i3} and r_{i4} the ratings student i would give TA 2 when that TA is identified as male and as female, respectively. Typically, the null hypotheses we test assert that for each i, some subset of $\{r_{ij}\}$ are equal. For assessing whether the identified gender of the TA affects SET, the null hypothesis is that for each i, $r_{i1} = r_{i2}$ (the rating the ith student would give TA 1 is the same, whether TA 1 is identified as male or female), and $r_{i3} = r_{i4}$ (the rating the ith student would give TA 2 is the same, whether TA 2 is identified as male or female). Different students might give different ratings under the same treatment condition (the null does not assert that $r_{ij} = r_{\ell j}$ for $i \neq \ell$), and the ith student might give different ratings to TA 1 and TA 2 (the null does not assert that $r_{i1} = r_{i3}$). The null hypothesis makes no assertion about the population distributions of $\{r_{i1}\}$ and $\{r_{i3}\}$, nor does it assert that $\{r_{ij}\}$ are a sample from some super-population.

For student i, we observe exactly one of $\{r_{i1}, r_{i2}, r_{i3}, r_{i4}\}$. If we observe r_{i1}, then—if the null hypothesis is true—we also know what r_{i2} is, and vice versa, but we do not know anything about r_{i3} or r_{i4}. Similarly, if we observe either r_{i3} or r_{i4} and the

null hypothesis is true, we know the value of both, but we do not know anything about r_{i1} or r_{i2}.

Consider the average SET (for any particular item) given by the N_2+N_4 students assigned to sections taught by an apparently female TA, minus the average SET given by the N_1+N_3 students assigned to sections taught by an apparently male TA. This is what MacNell et al. [2014] tabulate as their key result. If the perceived gender of the TA made no difference in how students rated the TA, we would expect the difference of averages to be close to zero.[5] How "surprising" is the observed difference in averages?

Consider the

$$\binom{N_1+N_2}{N_1} \times \binom{N_3+N_4}{N_3} \tag{4}$$

assignments that keep the same N_1+N_2 students in TA 1's sections (but might change which of those sections a student is in) and the same N_3+N_4 students in TA 2's sections. For each of those assignments, we know what $\{R_i\}_{i=1}^{N}$ would have been if the null hypothesis is true: each would be exactly the same as its observed value, since those assignments keep students in sections taught by the same TA. Hence, we can calculate the value that the test statistic would have had for each of those assignments.

Because all $\binom{N}{N_1 N_2 N_3 N_4}$ possible assignments of students to sections are equally likely, these $\binom{N_1+N_2}{N_1} \times \binom{N_3+N_4}{N_3}$ assignments in particular are also equally likely. The fraction of those assignments for which the value of the test statistic is at least as large (in absolute value) as the observed value of the test statistic is the p-value of the null hypothesis that students give the same rating (or none) to an TA, regardless

[5] We would expect it to be a least a little different from zero both because of the luck of the draw in assigning students to sections and because students might rate the two TAs differently, regardless of the TA's perceived gender, and the groups are not all the same size.

of the gender that TA appears to have.

This test is conditional on which of the students are assigned to each of the two TAs, but if we test at level no greater than α conditionally on the assignment, the unconditional level of the resulting test across all assignments is no greater than α, as shown above.

In principle, one could enumerate all the equally likely assignments and compute the value of the test statistic for each, to determine the (conditional) null distribution of the test statistic. In practice, there are prohibitively many assignments (for instance, there are $\binom{23}{11}\binom{24}{11} > 3.3 \times 10^{12}$ possible assignments of 47 students to the 4 TA-led sections that keep constant which students are assigned to each TA). Hence, we estimate p-values by simulation, drawing 10^5 equally likely assignments at random, with one exception, noted below. The distribution of the number of simulated assignments for which the test statistic is greater than or equal to its observed value is Binomial with n equal to the number of simulated assignments and p equal to the true p-value. Hence, the standard error of the estimated p-values is hence no larger than $(1/2)/\sqrt{10^5} \approx 0.0016$. Code for all our analyses is at https://github.com/kellieotto/SET-and-Gender-Bias. Results for the US data are in section 5.

4 The French Natural Experiment

In this section, we test hypotheses about relationships among SET, teaching effectiveness, grade expectations, and student and instructor gender. Our tests aggregate data within course sections, to match how SET are typically used in personnel deci-

sions. We use the average of Pearson correlations across strata as the test statistic,[6] which allows us to test both for differences in means (which can be written as correlations with a dummy variable) and for association with ordinal or quantitative variables.

In these analyses, individual i is a section of a course; the "treatment" is the instructor's gender, the average interim grade, or the average final exam score; and the "response" is the average SET or the average final exam score. Strata consist of all sections of a single course in a single year.

Our tests for overall effects stratify on the course subject, to account for systematic differences across departments: the hypothetical randomization shuffles characteristics among courses in a given department, but not across departments. We also perform tests separately in different departments, and in some cases separately by student gender.

4.1 SET and final exam scores

We test whether average SET scores and average final exam scores for course sections are associated (Table 2). The null hypothesis is that the pairing of average final grade and average SET for sections of a course each year is as if at random, independent across courses and across years. We test this hypothesis overall and separately by discipline, using the average Pearson correlation across strata, as described in section 3.1. If the null hypothesis were true, we would expect the test statistic to be close to zero. On the other hand, if SET do measure teaching effectiveness, we would expect average SET and average final exam score to be positively correlated

[6] As discussed above, we find p-values from the (nonparametric) permutation distribution, not from the theoretical distribution of the Pearson correlation under the parametric assumption of bivariate normality.

within courses within years, making the test statistic positive.

The numbers show that SET scores do not measure teaching effectiveness well, overall: the one-sided p-value for the hypothesis that the correlation is zero is 0.09. Separate tests by discipline find that for History, the association is positive and statistically significant (p-value of 0.01), while the other disciplines (Macroeconomics, Microeconomics, Political science and Sociology), the association is either negative or positive but not statistically significant (p-values 0.19, 0.55, 0.62, and 0.61 respectively).

Table 2: Average correlation between SET and final exam score, by subject

	strata	$\bar{\rho}$	p-value
Overall	26 (21)	0.04	0.09
History	5	0.16	0.01
Political Institutions	5	N/A	N/A
Macroeconomics	5	0.06	0.19
Microeconomics	5	-0.01	0.55
Political science	3	-0.03	0.62
Sociology	3	-0.02	0.61

Note: p-values are one-sided, since, if SET measured teaching effectiveness, mean SET should be positively associated with mean final exam scores. Correlations are computed for course-level averages of SET and final exam score within strata, then averaged across strata. Political Institutions is not reported, because the final exam was not graded anonymously. The five strata of Political Institutions are not included in the overall average, which is computed from the remaining 21 strata-level correlation coefficients.

4.2 SET and Instructor Gender

The second null hypothesis we test is that the pairing (by section) of instructor gender and SET is as if at random within courses each year, independently across years and courses. If gender does not affect SET, we would expect the correlation between average SET and instructor gender to be small in each course in each year. On the other hand, if students tend to rate instructors of one gender higher, we would

expect the average correlation to be large in absolute value. We find that average SET are significantly associated with instructor gender, with male instructors getting higher ratings (overall p-value 0.00). Male instructors get higher SET on average in every discipline (Table 3) with two-sided p-values ranging from 0.08 for History to 0.63 for Political Science.

Table 3: Average correlation between SET and instructor gender

	$\bar{\rho}$	p-value
Overall	0.09	0.00
History	0.11	0.08
Political institutions	0.11	0.10
Macroeconomics	0.10	0.16
Microeconomics	0.09	0.16
Political science	0.04	0.63
Sociology	0.08	0.34

Note: p-values are two-sided.

4.3 Instructor Gender and Learning Outcomes

Do men receive higher SET scores overall because they are better instructors? The third null hypothesis we test is that the pairing (by course) of instructor gender and average final exam score is as if at random within courses each year, independent across courses and across years. If this hypothesis is true, we would expect the average correlations to be small. If the effectiveness of instructors differs systematically by gender, we would expect average correlation to be large in absolute value. Table 4 shows that on the whole, students of male instructors perform worse on the final than students of female instructors, by an amount that is statistically significant (p-value 0.07 overall). In all disciplines, students of male instructors perform worse, but by amounts that are not statistically significant (p-values ranging from 0.22 for History to 0.70 for Political Science). This suggests that male instructors are not

noticeably more effective than female instructors, and perhaps are less effective: The statistically significant difference in SET scores for male and female instructors does not seem to reflect a difference in their teaching effectiveness.

Table 4: Average correlation between final exam scores and instructor gender

	$\bar{\rho}$	p-value
Overall	-0.06	0.07
History	-0.08	0.22
Macroeconomics	-0.06	0.37
Microeconomics	-0.06	0.37
Political science	-0.03	0.70
Sociology	-0.05	0.55

Note: p-values are two-sided. Negative values of $\bar{\rho}$ indicate that students of female instructors did better on average than students of male instructors.

4.4 Gender Interactions

Why do male instructors receive higher SET scores? Separate analyses by student gender shows that male students tend to give higher SET scores to male instructors (Table 5). These permutation tests confirm the results found by Boring [2015a]. Gender concordance is a good predictor of SET scores for men (p-value 0.00 overall). Male students give significantly higher SET scores to male instructors in History (p-value 0.01), Microeconomics (p-value 0.01), Macroeconomics (p-value 0.04), Political Science (p-value 0.06), and Political Institutions (p-value 0.08). Male students give higher SET scores to male instructors in Sociology as well, but the effect is not statistically significant (p-value 0.16).

The correlation between gender concordance and overall satisfaction scores for female students is also positive overall and weakly significant (p-value 0.09). The correlation is negative in some fields (History, Political Institutions, Macroeconomics,

Microeconomics and Sociology) and positive in only one field (Political Science), but in no case statistically significant (p-values range from 0.12 to 0.97).

Table 5: Average correlation between SET and gender concordance

	Male student		Female student	
	$\bar{\rho}$	p-value	$\bar{\rho}$	p-value
Overall	0.15	0.00	0.05	0.09
History	0.17	0.01	-0.03	0.60
Political institutions	0.12	0.08	-0.11	0.12
Macroeconomics	0.14	0.04	-0.05	0.49
Microeconomics	0.18	0.01	-0.00	0.97
Political science	0.17	0.06	0.04	0.64
Sociology	0.12	0.16	-0.03	0.76

Note: p-values are two-sided.

Do male instructors receive higher SET scores from male students because their teaching styles match male students' learning styles? If so, we would expect male students of male instructors to perform better on the final exam. However, they do not (Table 6). If anything, male students of male instructors perform worse overall on the final exam (the correlation is negative but not statistically significant, with a p-value 0.75). In History, the amount by which male students of male instructors do worse on the final is significant (p-value 0.03): male History students give significantly higher SET scores to male instructors, despite the fact that they seem to learn more from female instructors. SET do not appear to measure teaching effectiveness, at least not primarily.

4.5 SET and grade expectations

The next null hypothesis we test is that the pairing by course of average SET scores with average interim grades is as if at random. Because interim grades may set student grade expectations, if students give higher SET in courses where they expect

Table 6: Average correlation between student performance and gender concordance

	Male student		Female student	
	$\bar{\rho}$	p-value	$\bar{\rho}$	p-value
Overall	-0.01	0.75	0.06	0.07
History	-0.15	0.03	-0.02	0.74
Macroeconomics	0.04	0.60	0.11	0.10
Microeconomics	0.02	0.80	0.07	0.29
Political science	0.08	0.37	0.11	0.23
Sociology	0.01	0.94	0.06	0.47

Note: p-values are two-sided.

higher grades, the association should be positive. Indeed, the association is positive and generally highly statistically significant (Table 7). Political institutions is the only discipline for which the average correlation between interim grades and SET scores is negative, but the correlation is not significant (p-value 0.61). The estimated p-values for all other courses are between 0.0 and 0.03. The average correlations are especially high in History (0.32) and Sociology (0.24).

Table 7: Average correlation between SET and interim grades

	$\bar{\rho}$	p-value
Overall	0.16	0.00
History	0.32	0.00
Political institutions	-0.02	0.61
Macroeconomics	0.15	0.01
Microeconomics	0.13	0.03
Political science	0.17	0.02
Sociology	0.24	0.00

Note: p-values are one-sided.

In summary, the average correlation between SET and final exam grades (at the level of class sections) is positive, but only weakly significant overall and not significant for most disciplines. However, the average correlation between SET and grade expectations (at the level of class sections) is positive and significant overall

and across most disciplines. The average correlation between instructor gender and SET is statistically significant—male instructors get higher SET—but if anything, students of male instructors do worse on final exams than students of female instructors. Male students tend to give male instructors higher SET, even though they might be learning less than they do from female instructors. We conclude that SET are influenced more by instructor gender and student grade expectations than by teaching effectiveness.

5 The US Randomized Experiment

The previous section suggests that SET have little connection to teaching effectiveness, but the natural experiment does not allow us to control for differences in teaching styles across instructors. MacNell et al. [2014] does. As discussed above, MacNell et al. [2014] collected SET from an online course in which 43 students were randomly assigned to four[7] discussion groups, each taught by one of two TAs, one male and one female. The TAs gave similar feedback to students, returned assignments at exactly the same time, etc.

Biases in student ratings are revealed by differences in ratings each TA received when that TA is identified to the students as male versus as female. 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

[7]As discussed above, there were six sections in all, of which two were taught by the professor and four were taught by TAs.

gender of the assistant instructor." MacNell et al. [2014] used parametric tests whose assumptions did not match their experimental design; part of our contribution is to show that their data admit a more rigorous analysis using permutation tests that honor the underlying randomization and that avoid parametric assumptions about SET. The new analysis supports their overall conclusions, in some cases substantially more strongly than the original analysis (for instance, p-values of 0.01 versus 0.19 for promptness and fairness). In other cases, the original parametric tests overstated the evidence (for instance, a p-value of 0.29 versus 0.04 for knowledgeability).

We use permutation tests as described above in section 3. Individual i is a student; the treatment is the combination of the TA's identity and the TA's apparent gender (there are $K = 4$ treatments). The null hypothesis is that each student would give a TA the same SET score, whether that TA is apparently male or apparently female. A student might give the two TAs different scores, and different students might give different scores to the same TA.

Because of how the experimental randomization was performed, all allocations of students to TA sections that preserve the number of students in each section are equally likely, including allocations that keep the same students assigned to each actual TA constant.

To test whether there is a systematic difference in how students rate apparently male and apparently female TAs, we use the difference in pooled means as our test statistic: We pool the SET for both instructors when they are identified as female and take the mean, pool the SET for both instructors when they are identified as male and take the mean, then subtract the second mean from the first mean (Table 8). This is what MacNell et al. [2014] report as their main result.

As described above, the randomization is stratified and conditions on the set of students allocated to each TA, because, under the null hypothesis, we then know what

SET students would have given for each possible allocation, completely specifying the null distribution of the test statistic. The randomization includes the nonresponders, who are omitted from the averages of the group they are assigned to.

We also perform tests involving the association of concordance of student and apparent TA gender, (Table 9) and SET and concordance of student and actual TA gender (Table 10) using the pooled difference in means as the test statistic. We test the association between grades and actual TA gender (Table 11) using the average Pearson correlation across strata as the test statistic. We find the p-values from the stratified permutation distribution of the test statistic, avoiding parametric assumptions.

5.1 SET and Perceived Instructor Gender

The first hypothesis we test is that students would rate a given TA the same, whether the student thinks the TA is female or male. A positive value of the test statistic means that students give higher SET on average to apparently male instructors. There is weak evidence that the overall SET score depends on the perceived gender (p-value 0.12). The evidence is stronger for several other items students rated: fairness (p-value 0.01), promptness (p-value 0.01), giving praise (p-value 0.01), enthusiasm (p-value 0.06), communication (p-value 0.07), professionalism (p-value 0.06), respect (p-value 0.07), and caring (p-value 0.09). For seven items, the nonparametric permutation p-values are smaller than the parametric p-values reported by MacNell et al. [2014]. Items for which the permutation p-values were greater than 0.10 include clarity, consistency, feedback, helpfulness, responsiveness, and knowledgeability. SET were on a 5-point scale, so a difference in means of 0.80, observed in student ratings of the promptness with which assignments were returned, is 16%

of the full scale—an enormous difference. Since assignments were returned at exactly the same time in all four sections of the class, this seriously impugns the ability of SET to measure even putatively objective characteristics of teaching.

Table 8: Mean ratings and reported instructor gender (male minus female)

	difference in means	nonparametric p-value	MacNell et al. p-value
Overall	0.47	0.12	0.128
Professional	0.61	0.07	0.124
Respectful	0.61	0.06	0.124
Caring	0.52	0.10	0.071
Enthusiastic	0.57	0.06	0.112
Communicate	0.57	0.07	NA
Helpful	0.46	0.17	0.049
Feedback	0.47	0.16	0.054
Prompt	0.80	0.01	0.191
Consistent	0.46	0.21	0.045
Fair	0.76	0.01	0.188
Responsive	0.22	0.48	0.013
Praise	0.67	0.01	0.153
Knowledge	0.35	0.29	0.038
Clear	0.41	0.29	NA

Note: p-values are two-sided.

We also conducted separate tests by student gender. In contrast to our findings for the French data, where male students rated male instructors higher, in the MacNell et al. [2014] experiment, perceived male instructors received significantly higher evaluation scores because female students rated the perceived male instructors higher (Table 9). Male students rated the perceived male instructor significantly (though weakly) higher on only one criterion: fairness (p-value 0.09). Female students, however, rated the perceived male instructor higher on overall satisfaction (p-value 0.11) and most teaching dimensions: praise (p-value 0.01), enthusiasm (p-value 0.05), caring (p-value 0.05), fairness (p-value 0.05), respectfulness (p-value 0.12), communication (p-value 0.10), professionalism (p-value 0.12), and feedback (p-value 0.10).

Female students rate (perceived) female instructors lower on helpfulness, promptness, consistency, responsiveness, knowledge, and clarity, although the differences are not statistically significant.

Table 9: SET and reported instructor gender (male minus female)

	Male students		Female students	
	difference in means	p-value	difference in means	p-value
Overall	0.17	0.82	0.79	0.11
Professional	0.42	0.55	0.82	0.12
Respectful	0.42	0.55	0.82	0.12
Caring	0.04	1.00	0.96	0.05
Enthusiastic	0.17	0.83	0.96	0.05
Communicate	0.25	0.68	0.87	0.10
Helpful	0.46	0.43	0.51	0.35
Feedback	0.08	1.00	0.88	0.10
Prompt	0.71	0.15	0.86	0.13
Consistent	0.17	0.85	0.77	0.17
Fair	0.75	0.09	0.88	0.04
Responsive	0.38	0.54	0.06	1.00
Praise	0.58	0.29	0.81	0.01
Knowledge	0.17	0.84	0.54	0.21
Clear	0.13	0.85	0.67	0.29

Note: p-values are two-sided.

Students of both genders rated the apparently male instructor higher on all dimensions, by an amount that often was statistically significant for female students (Table 9). However, students rated the actual male instructor higher on some dimensions and lower on others, by amounts that generally were not statistically significant (Table 10). The exceptions were praise (p-value 0.02) and responsiveness (p-value 0.05), where female students tended to rate the actual female instructor significantly higher.

Students of the actual male instructor performed worse in the course on average, by an amount that was statistically significant (Table 11). The difference in student performance by perceived gender of the instructor is not statistically significant.

Table 10: SET and actual instructor gender (male minus female)

	Male students		Female students	
	difference in means	p-value	difference in means	p-value
Overall	-0.13	0.61	-0.29	0.48
Professional	0.15	0.96	-0.09	0.73
Respectful	0.15	0.96	-0.09	0.73
Caring	-0.22	0.52	-0.07	0.75
Enthusiastic	-0.13	0.62	-0.44	0.29
Communicate	-0.02	0.80	-0.18	0.61
Helpful	0.03	0.89	0.26	0.71
Feedback	-0.24	0.48	-0.41	0.36
Prompt	-0.09	0.69	-0.33	0.44
Consistent	0.12	0.97	-0.40	0.35
Fair	-0.06	0.71	-0.59	0.12
Responsive	-0.13	0.64	-0.68	0.05
Praise	0.02	0.86	-0.60	0.02
Knowledge	0.22	0.83	-0.44	0.17
Clear	-0.26	0.49	-0.98	0.07

Note: p-values are two-sided.

Table 11: Mean grade and instructor gender (male minus female)

	difference in means	p-value
Perceived	1.76	0.54
Actual	-6.81	0.02

Note: p-values are two-sided.

These results suggest that students rate instructors more on the basis of the instructor's perceived gender than on the basis of the instructor's effectiveness. Students of the TA who is actually female did substantially better in the course, but students rated apparently male TAs higher.

6 Multiplicity

We did not adjust the p-values reported above for multiplicity. We performed a total of approximately 50 tests on the French data, of which we consider four to be our primary results:

1FR lack of association between SET and final exam scores (a negative result, so multiplicity is not an issue)

2FR lack of association between instructor gender and final exam scores (a negative result, so multiplicity is not an issue)

3FR association between SET and instructor gender

4FR association between SET and interim grades

Bonferroni's adjustment for these four tests would leave the last two associations highly significant, with adjusted p-values less than 0.01.

We performed a total of 77 tests on the US data. We consider the three primary null hypotheses to be

1US perceived instructor gender plays no role in SET

2US male students rate perceived male and female instructors the same

3US female students rate perceived male and female instructors the same

To account for multiplicity, we tested these three "omnibus" hypotheses using the nonparametric combination of tests (NPC) method with Fisher's combining function [Pesarin and Salmaso, 2010, Chapter 4] to summarize the 15 dimensions of teaching into a single test statistic that measures how "surprising" the 15 observed differences would be for each of the three null hypotheses. In 10^5 replications, the empirical p-values for these three omnibus hypotheses were 0 (99% confidence interval $[0.0, 5.3 \times 10^{-5}]$) 0.464 (99% confidence interval $[0.460, 0.468]$), and 0 (99% confidence interval $[0.0, 5.3 \times 10^{-5}]$), respectively. (The confidence bounds were obtained by inverting Binomial hypothesis tests.) Thus, we reject hypotheses 1US and 3US.

We made no attempt to optimize the tests to have power against the alternatives considered. For instance, with the US data, the test statistic grouped the two identified-as-female sections and the two identified-as-male conditions, in keeping with how MacNell et al. [2014] tabulated their results, rather than using each TA as his or her own control (although the randomization keeps the two strata intact). Given the relatively small number of students in the US experiment, it is remarkable that *any* of the p-values is small, much less that the p-values for the omnibus tests are effectively zero.

7 Code and Data

Jupyter (`http://jupyter.org/`) notebooks containing our analyses are at `https://github.com/kellieotto/SET-and-Gender-Bias`; they rely on the `permute` Python library (`https://pypi.python.org/pypi/permute/`). The US data are available at `http://n2t.net/ark:/b6078/d1mw2k`. French privacy law prohibits publishing the French data.

8 Discussion

8.1 Other studies

To our knowledge, only two experiments have controlled for teaching style in their designs: Arbuckle and Williams [2003] and MacNell et al. [2014]. In both experiments, students generally gave higher SET when they *thought* the instructor was male, regardless of the actual gender of the instructor. Both experiments found that systematic differences in SET by instructor gender reflect gender bias rather than a match of teaching style and student learning style or a difference in actual teaching effectiveness.

Arbuckle and Williams [2003] showed a group of 352 students "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]. All students saw the same stick figure and heard the same voice, so differences in SET could be attributed to the age and gender the students were *told* the instructor had. When students were told the instructor was young and male, students rated the instructor higher than for the other three combinations, especially on "enthusiasm," "showed interest in subject," and "using a meaningful voice tone."

Instructor race is also associated with SET. In the US, SET of instructors of color appear to be biased downwards: minority instructors tend to receive significantly lower SET scores compared to white (male) instructors [Merritt, 2008].[8] Age, [Arbuckle and Williams, 2003], charisma [Shevlin et al., 2000], and physical attractiveness [Riniolo et al., 2006, Hamermesh and Parker, 2005] are also associated

[8]French law does not allow the use of race-related variables in data sets. We were thus unable to test for racial biases in SET using the French data.

with SET. Other factors generally not in the instructor's control that may affect SET scores include class time, class size, mathematical or technical content, and the physical classroom environment [Hill and Epps, 2010].

Many studies cast doubt on the validity of SET as a measure of teaching effectiveness (see Johnson [2003, Chapters 3–5] for a review and analysis, Pounder [2007] for a review, and Galbraith et al. [2012], Carrell and West [2010] for exemplars). Some studies find that gender and SET are not significantly associated [Bennett, 1982, Centra and Gaubatz, 2000, Elmore and LaPointe, 1974] and that SET are valid and reliable measures of teaching effectiveness [Benton and Cashin, 2012, Centra, 1977].[9] The contradictions among conclusions suggests that if SET are ever valid, they are not valid in general: universities should not assume that SET are broadly valid at their institution, valid in any particular department, or valid for any particular course. Given the many sources of bias in SET and the variability in magnitude of the bias by topic, item, student gender, etc., as a practical matter it is impossible to adjust for biases to make SET a valid, useful measure of teaching effectiveness.

8.2 Summary

We used permutation tests to examine data collected by Boring [2015a] and MacNell et al. [2014], both of which find that gender biases prevent SET from measuring teaching effectiveness accurately and fairly. SET are more strongly related to instructor's perceived gender and to students' grade expectations than they are to learning, as measured by performance on anonymously graded, uniform final exams. The extent and direction of gender biases depend on context, so it is impossible to adjust for such biases to level the playing field. While the French university data show a

[9]Some authors who claim that SET are valid have a financial interest in developing SET instruments and conducting SET.

positive male student bias for male instructors, the experimental US setting suggests a positive female student bias for male instructors. The biases in the French university data vary by course topic; the biases in the US data vary by item. We would also expect the bias to depend on class size, format, level, physical characteristics of the classroom, instructor ethnicity and a host of other variables.

We do not claim that there is *no* connection between SET and student performance. However, the observed association is sometimes positive and sometimes negative, and in general is not statistically significant—in contrast to the statistically significant strong associations between SET and grade expectations and between SET and instructor gender. SET appear to measure student satisfaction and grade expectations more than they measure teaching effectiveness [Stark and Freishtat, 2014, Johnson, 2003]. While student satisfaction may *contribute* to teaching effectiveness, it is not itself teaching effectiveness. Students may be satisfied or dissatisfied with courses for reasons unrelated to learning outcomes—and not in the instructor's control (e.g., the instructor's gender).

In the US, SET have two primary uses: instructional improvement and personnel decisions, including hiring, firing, and promoting instructors. We recommend caution in the first use, and discontinuing the second use, given the strong student biases that influence SET, even on "objective" items such as how promptly instructors return assignments.[10]

[10]In 2009, the French Ministry of Higher Education and Research upheld 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 (c.f. Boring [2015b]).

8.3 Conclusion

In two very different universities and in a broad range of course topics, SET measure students' gender biases better than they measure the instructor's teaching effectiveness. Overall, SET disadvantage female instructors. There is no evidence that this is the exception rather than the rule. Hence, the onus should be on universities that rely on SET for employment decisions to provide convincing affirmative evidence that such reliance does not have disparate impact on women, under-represented minorities, or other protected groups. Because the bias varies by course and institution, affirmative evidence needs to be specific to a given course in a given department in a given university. Absent such specific evidence, SET should not be used for personnel decisions.

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

V. E. Johnson. *Grade Inflation: A Crisis in College Education*. Springer-Verlag, New York, 2003.

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. Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472, 1990.

F. Pesarin and L. Salmaso. *Permutation Tests for Complex Data: Theory, Applications and Software*. Wiley, New York, 2010.

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. doi: 10.14293/S2199-1006.1.-.AOFRQA.v1. URL https://www.scienceopen.com/document/vid/42e6aae5-246b-4900-8015-dc99b467b6e4.