scienceOPENresearch SOR-EDU

Student evaluations of teaching (mostly) do not measure teaching effectiveness

Anne Boring[1,2], Kellie Ottoboni[3], and Philip B. Stark*[3]

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.

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.

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

Gloria Steinem

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 [2,3] 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 [4, Chapters 3–5].

Gender matters too. Boring [5] 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. [6] finds that when students think an instructor is female, students rate the instructor lower on every aspect of teaching, including putatively objective measures such as the timeliness with which instructors return assignments.

[1] Centra and Gaubatz [1, 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."

Here, we apply nonparametric permutation tests to data from Boring [5] and MacNell et al. [6] 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, and logistic regression) and methods such as *t*-tests and ANOVA generally require. The null hypotheses for our tests are that some characteristics – e.g., instructor gender – amount 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 [7] 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.

DATA

US randomized experiment

These data, described in detail by MacNell et al. [6], 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 sections, 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. Forty-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] while the grade data do not. The SET data are not linked to the grade data.

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 estimates, and no model.

[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.

[4] The final exam in political institutions is oral and hence not graded anonymously.

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. [6] does. As discussed above, MacNell et al. [6] 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.

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. [6] 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." MacNell et al. [6] 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

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

supports their overall conclusions, in some cases substantially more strongly than the original analysis (for instance, a p-value of 0.01 vs. 0.19 for promptness and fairness). In other cases, the original parametric tests overstated the evidence (for instance, a p-value of 0.29 vs. 0.04 for knowledgeability).

We use permutation tests as described above in "Methods" section. 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 and when they are identified as male and take the mean, then subtract the second mean from the first mean (Table 8), as reported by MacNell et al. [6] as their main result.

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

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

p-values are two-sided.

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

	Male students		Female students	
	Difference in means	p	Difference in means	p
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

p-values are two-sided.

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 0.07), respect (p-value 0.06), and caring (p-value 0.10). For seven items, the nonparametric permutation p-values are smaller than the parametric p-values reported by MacNell et al. [6]. 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 20% of the full range – 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.

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.

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

	Male students		Female students	
	Difference in means	*p*	Difference in means	*p*
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

p-values are two-sided.

[6] 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 0.04), 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. 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 11. Mean grade and instructor gender (male minus female).

	Difference in means	*p*
Perceived	1.76	0.54
Actual	−6.81	0.02

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.

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 [10, 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 estimated *p*-values for these three omnibus hypotheses were 0 (99% confidence interval [0.0, 5.3×10^{-5}]), 0.464 (99% confidence interval [0.460, 0.468]), and 0 (99% confidence interval [0.0, 5.3×10^{-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. [6] 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.

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.

DISCUSSION

Other studies

To our knowledge, only two experiments have controlled for teaching style in their designs: Arbuckle and Williams [11] and MacNell et al. [6]. 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 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')" [11, 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 downward: minority instructors tend to receive significantly lower SET scores compared to white (male) instructors [12].[8] Age, [11], charisma [13], and physical attractiveness [14,15] are also associated 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 [16], and the physical classroom environment [17].

Many studies cast doubt on the validity of SET as a measure of teaching effectiveness (see Johnson [4, Chapters 3-5] for a review and analysis, Pounder [18] for a review, and Galbraith et al. [19], Carrell and West [2] for exemplars). Some studies find that gender and SET are not significantly associated [1,20,21]. Those studies generally address a different, namely, whether men and women receive similar SET. That does not control for teaching effectiveness, effort, or other variables. The more relevant question is whether women would receive higher scores for doing the same thing had they been male, and whether men would receive lower scores for doing the same thing had they been female. Our analysis of the US data shows that is true. Our analysis of the French data shows that, on average, less effective male instructors receive higher SET than more effective female instructors.

Some studies find that SET are valid and reliable measures of teaching effectiveness [22,23].[9] The contradictions among conclusions suggest 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, and so on, as a practical matter it is impossible to adjust for biases to make SET a valid, useful measure of teaching effectiveness.

Summary

We used permutation tests to examine data collected by Boring [5] and MacNell et al. [6], 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 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 [4,7]. 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 [6].[10]

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

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

[10] In 2009, the French Ministry of Higher Education and Research upheld a 1997 decision of the French State Council

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, underrepresented 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.

FUNDING

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 612413, for the EGERA (Effective Gender Equality in Research and the Academia) European project.

REFERENCES

[1] Centra JA, Gaubatz NB. Is there gender bias in student evaluations of teaching? J High Educ. 2000;71(1):17–33. doi:10.2307/2649280

[2] Carrell SE, West JE. Does professor quality matter? Evidence from random assignment of students to professors. J Polit Econ. 2010;118(3):409–432. ISSN 0022-3808. doi: 10.1086/653808

[3] Braga M, Paccagnella M, Pellizzari M. Evaluating students evaluations of professors. Econ Educ Rev. 2014;41:71–88.

[4] Johnson VE. Grade inflation: a crisis in college education. New York: Springer-Verlag; 2003.

[5] Boring A. Gender biases in student evaluations of teachers. Document de travail OFCE 13. Paris, France: OFCE; 2015a.

[6] MacNell L, Driscoll A, Hunt AN. What's in a name? Exposing gender bias in student ratings of teaching. Innovat High Educ. 2015;40(4):291–303.

[7] Stark PB, Freishtat R. An evaluation of course evaluations. Sci Open Res. 2014:1–7. doi:10.14293/S2199-1006.1.-.AOFRQA.v1

[8] Marsh HW, Roche LA. Making students' evaluations of teaching effectiveness effective. Am Psychol. 1997;52(11):1187–1197.

[9] Neyman J, Dabrowska DM, Speed TP. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat Sci. 1990;5(4):465–472.

[10] Pesarin F, Salmaso L. Permutation tests for complex data: theory, applications and software. New York: Wiley; 2010.

[11] Arbuckle J, Williams BD. Students' perceptions of expressiveness: age and gender effects on teacher evaluations. Sex Roles. 2003;49:507–516.

[12] Merritt DJ. Bias, the brain, and student evaluations of teaching. St. John's Law Rev. 2008;81(1):235–288.

[13] Shevlin M, Banyard P, Davies M, Griffiths M. The validity of student evaluation of teaching in higher education: love me, love my lectures? Assess Eval High Educ. 2000;25(4):397–405.

[14] Riniolo TC, Johnson KC, Sherman TR, Misso JA. Hot or not: do professors perceived as physically attractive receive higher student evaluations? J Gen Psychol. 2006;133(1):19–35. ISSN 0022-1309. doi:10.3200/GENP.133.1.19-35

[15] Hamermesh DS, Parker A. Beauty in the classroom: instructors pulchritude and putative pedagogical productivity. Econ Educ Rev 2005;24(4):369–376.

[16] Royal KD, Stockdale MR. Are Teacher Course Evaluations Biased Against Faculty That Teach Quantitative Methods Courses? Int J Higher Educ 2015;4(1):217–224. ISSN 1927-6044. E-ISSN 1927-6052.

[17] Hill MC, Epps KK. The impact of physical classroom environment on student satisfaction and student evaluation of teaching in the university environment. Acad Educ Leader J. 2010;14(4):65–79.

[18] Pounder JS. Is student evaluation of teaching worthwhile?: An analytical framework for answering the question. Qual Assur Educ. 2007;15(2):178–191. ISSN 0968-4883. doi:10.1108/09684880710748938

[19] Galbraith CS, Merrill GB, Kline DM. Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and Bayesian analyses. Res High Educ. 2012;53(3):353–374.

[20] Bennett SK. Student perceptions of and expectations for male and female instructors: evidence relating to the question of gender bias in teaching evaluation. J Educ Psychol. 1982;74(2):170–179.

[21] Elmore PB, LaPointe KA. Effects of teacher sex and student sex on the evaluation of college instructors. J Educ Psychol. 1974;66(3):386–389.

[22] Benton SL, Cashin WE. Student ratings of teaching: a summary of research and literature. IDEA Paper 50. Manhattan, KS: The IDEA Center; 2012.

[23] Centra JA. Student ratings of instruction and their relationship to student learning. Am Educ Res J. 1977;14(1):17–24.

[24] Boring A. Can students evaluate teaching quality objectively? Le blog de l'ofce. OFCE; 2015b. Available from http://www.ofce.sciences-po.fr/blog/can-students-evaluate-teaching-quality-objectively/. [cited 24 February 2015].

COMPETING INTERESTS

The authors declare no competing interests.

PUBLISHING NOTES

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 [24]).