Expert Report on Student Evaluations of Teaching (Faculty Course Surveys)

Prepared for
The Ryerson Faculty Association
and
The Ontario Confederation of University Faculty Associations

Philip B. Stark, Ph.D.

10 October 2016

I. QUALIFICATIONS AND BACKGROUND

1. I am Professor of Statistics, Associate Dean of Mathematical and Physical Sciences, and Director of the Statistical Computing Facility at the University of California, Berkeley, where I am also a faculty member in the Graduate Program in Computational Data Science and Engineering; a co-investigator at the Berkeley Institute for Data Science; principal investigator of the Consortium for Data Analytics in Risk; director of Berkeley Open Source Food; and affiliated faculty of the Simons Institute for the Theory of Computing, the Theoretical Astrophysics Center, and the Berkeley Food Institute. Previously, I was Chair of the Department of Statistics. I have also had campus-wide responsibilities regarding educational technology, including technology involved in teaching evaluations.

2. I have published more than one hundred and fifty articles and books. I have served on the editorial boards of archival journals in physical science, Applied Mathematics, Computer Science, and Statistics. I currently serve on four editorial boards. I have lectured at universities, professional societies, and government agencies in twenty-five countries. I was a Presidential Young Investigator, a Miller Research Professor, and a Velux/Villum Foundation Visiting Professor of Theoretical Computer Science. I received the U.C. Berkeley Chancellor's Award for Research in the Public Interest and the Leamer-Rosenthal Prize for Open Social Science. I am a member of the Institute for Mathematical Statistics and the Bernoulli Society; and I am a Fellow of the American Statistical Association, the Institute of Physics, and the Royal Astronomical Society. I am professionally accredited as a statistician by the American Statistical Association and as a physicist by the Institute of Physics.

3. I have consulted for many U.S. government agencies, including the U.S. Department of Justice, the U.S. Department of Agriculture, the U.S. Department of Commerce, the U.S. Department of Housing and Urban Development, the U.S. Department of Veterans Affairs, the Federal Trade Commission, the California Secretary of State, the California Attorney General, the California Highway Patrol, the Colorado Secretary of State, and the Illinois State Attorney. I serve on the Advisory Board of the U.S. Election Assistance Commission.

4. I have testified before the U.S. House of Representatives Subcommittee on the Census; the State of California Senate Committee on Elections, Reapportionment and Constitutional Amendments; the State of California Assembly Committee on Elections and Redistricting; and the State of California Senate Committee on Natural Resources.

5. I have been an expert witness or non-testifying expert in state and federal cases, for plaintiffs and for defendants, in criminal matters and a range of civil matters, including, *inter alia*: truth in advertising, antitrust, construction defects, consumer class actions, credit risk, disaster relief, election contests, employment discrimination, environmental protection, equal protection, fairness in lending, federal legislation, First Amendment, import restrictions, insurance, intellectual property, jury selection, mortgage-backed securities, natural resources, product liability class actions, *qui tam*, risk assessment, toxic tort class actions, trade secrets, and wage and hour class actions.

6. I have been qualified as an expert on statistics in U.S. federal courts, including the Central District of California, the District of Maryland, the Southern District of New York, and the Eastern District of Pennsylvania. I have also been qualified as an expert in statistics in state courts, and I have testified as an expert in binding arbitrations.

7. I have used statistics to address a wide range of questions in many fields.[1]

8. I developed statistical approaches to auditing elections ("risk-limiting audits") that have been incorporated into statutes in California (AB 2023, SB 360, AB 44) and Colorado (C.R.S. 1-7-515). Statistical approaches to data reduction that I developed are used by the Danish Ørsted satellite and by the Global Oscillations Network Group international network of solar telescopes.

9. Since 1988, I have taught statistics at the University of California, Berkeley, one of the top two statistics departments in the world (see, e.g., QS World University Rankings, 2014) and the nation (US News and World Reports, 2014). I teach statistics regularly at

[1] For example, I have used statistics to analyze the Big Bang, the interior structure of the Earth and Sun, the risk of large earthquakes, the reliability of clinical trials, the accuracy of election results, the accuracy of the U.S. Census, the risk of consumer credit default, the causes of geriatric hearing loss, the effectiveness of water treatment, the fragility of ecological food webs, risks to protected species, the effectiveness of Internet content filters, and the reliability models of climate, among other things.

the undergraduate and graduate levels. I have created five new statistics courses at U.C. Berkeley. I developed and taught U.C. Berkeley's first online course in any subject (Statistics W21, subsequently approved for credit throughout the ten campuses of the University of California system). I also developed and co-taught online statistics courses to over 52,000 students, based on an online textbook and other pedagogical materials I wrote and programmed.

10. I have professional experience in the evaluation of teaching and the administration of student evaluations. I have published two research papers on student evaluations, their use and misuse, and their biases; one of the papers has been downloaded more than 32,000 times. I have testified in union grievances and binding arbitration on the use of student evaluations of teaching (SET) for employment decisions. My work on SET has received media attention in academic journals and the popular press, from *National Public Radio* and *The Chronicle of Higher Education* to *Seventeen*.

11. I designed statistical tests of online SET deployed by U.C. Berkeley in fall, 2012. I have also conducted experiments on the association between student ratings of instructor effectiveness and student enjoyment.

12. While chair of the Department of Statistics, I made major changes to how teaching was evaluated for the purpose of employment decisions, including introducing teaching portfolios and peer observation before milestone reviews, such as tenure and promotion to full professor. Those changes were adopted by the Dean of Mathematical and Physical Sciences as a model for improving the evaluation of teaching. I was twice invited to address all academic U.C. Berkeley administrators, including deans and department chairs, on the use of SET, biases in SET, and improper reliance on SET. I have also been invited to present my work on SET to the deans and department chairs within the College of Letters and Sciences and, separately, within the Division of Mathematical Sciences.

13. I have given professional presentations about the use and misuse of SET at academic conferences, departmental seminars, university-wide seminars, and workshops for university administrators, including lectures at The University of Pennsylvania, The University of California Santa Cruz, Colorado State University, The University of San

Francisco, the Center for Studies in Higher Education at The University of California, Berkeley, and the National Center for the Study of Collective Bargaining in Higher Education and the Professions at Hunter College. Next month, I am scheduled to make a presentation on SET at the University of California, San Diego.

14. My role as Associate Dean includes adjudicating employment grievances brought by represented academic staff. I have adjudicated an employment grievance in which SET played a central role.

15. In my role as Associate Dean, I am responsible for developing methods for evaluating teaching and promoting excellence in teaching that are less subject to the biases that pervade SET.

16. My current *curriculum vitae* is attached.

II. SUMMARY OF OPINIONS

17. I understand that student evaluations of teaching (SET) are called "Faculty Course Surveys" (FCSs) at Ryerson University. I use the term SET below.

18. There is a large literature on SET. The best evidence about the connection between SET and other variables comes from experiments that assign students at random to sections of courses, in a manner similar to clinical trials. By comparing student performance and SET across sections, one can establish the extent to which SET measure the effectiveness of instruction, or are influenced by other factors, such as the gender of the instructor.

19. Such experiments, along with other large, multi-section studies, generally find weak or negative association between SET and instructor effectiveness, measured by performance on uniformly graded final exams or performance in follow-on courses (Carrell and West, 2010; Boring et al., 2016; Braga et al., 2014; Johnson, 2003, especially Ch.5; MacNell et al., 2015; Uttl et al., 2016). The best evidence suggests that SET are neither reliable nor valid, even when the survey response rate is nearly perfect.

20. However, there are papers that argue that SET are reliable and valid. The studies I have seen are not convincing; in particular, none rises to the level of rigor of the randomized experiments and "natural experiments" cited above. Generally, they lack appropriate controls, do not use randomization, use inappropriate statistical tests, and conflate statistical significance with effect size.

21. Suppose that SET were in some instances reliable and valid. Because in many circumstances SET are biased, as described below, SET should not be presumed to be valid, reliable, or fair in any given course, department, or university, absent affirmative evidence of reliability, validity, and unbiasedness *in that time and place.*

22. There is substantial evidence that SET have large biases. Sources of bias include students' grade expectations (e.g., Boring et al., 2016; Marsh and Cooper, 1980; Vasta and Sarmiento, 1979); the nature of the course material (for instance, instructors who teach courses with mathematical content tend to get lower ratings, Uttl et al., 2013), the level of the course and whether the course is required (e.g., Marsh and Roche, 1997), the course format (Lake, 2001), the size of the course (Bedard and Kuhn, 2005), instructor gender (Arbuckle and Williams, 2003; Basow et al., 2013; Bianchini et al., 2013; Boring, 2015; Boring et al., 2016; MacNell et al., 2015), instructor age (Arbuckle and Williams, 2003; Bianchini et al., 2013), instructor attractiveness (Ambady and Rosenthal, 1993; Wolbring and Riordan, 2016), instructor expressiveness (Ambady and Rosenthal, 1993; Williams and Ceci, 1997), instructor race (Archibeque, 2014, and citations therein; Basow et al., 2013), whether the instructor speaks with an accent or is a native speaker (Subtirelu, 2015), the physical condition of the classroom (Hill and Epps, 2010), and so on. Many of these factors are protected characteristics under employment law: relying on student evaluations may have disparate impact on protected groups. Other factors may not be in the control of the instructor.

23. The biases can be so large that more effective teachers get lower SET than less effective teachers (Boring et al., 2016). There is evidence that the biases vary by discipline (and other variables, including student gender), making it essentially

impossible to adjust SET for bias to obtain a fair and meaningful measure of teaching quality or effectiveness (Boring et al., 2016).

24. "Omnibus" items on SET, such as questions about effectiveness, are particularly subject to bias (Worthington, 2002). However, even putatively objective items, such as whether assignments are returned promptly, are subject to large biases (Boring et al., 2016; MacNell et al., 2015).

25. In short, it is possible that SET are reliable measures of students' experiences (Did the student enjoy the class? Could the student read the instructor's handwriting?) but SET are generally unreliable, biased, and invalid measures of items that require judgment (Was the instructor/course effective? Was the instructor professional? Was the instructor fair? Was the course material well organized?) or accurate memory (for instance, there is evidence that students do not accurately report the number of hours per week they typically spend working on a course, nor whether instructors are generally available outside class).

26. Even if an SET item measured what it purports to measure, it would be statistically inappropriate and misleading to average SET scores and to compare average scores across courses, instructors, disciplines, and so on (Stark and Freishtat, 2014). In part this is because many variables in question are "ordinal categorical" variables, rather than quantitative, linear variables (Stark and Freishtat, 2014).

27. A *categorical variable* is a variable whose possible values are *labels* or *categories*, such as "blue, green, red, yellow." An *ordinal categorical variable* is a categorical variable whose possible values have a natural order, for instance, "strongly disagree, disagree, neither agree nor disagree, agree, strongly agree" (which are ordered by strength of agreement) and "cold, cool, warm, hot" (which are ordered by temperature).

28. While it is common to replace the category names with numbers, for instance, using "1" to signify "strongly disagree" and "5" to signify "strongly agree," the numbers themselves are not quantities, just new labels. They are codes that happen to be

numerical. The actual magnitudes of the numbers do not mean anything. The labels are arbitrary.

29. Averaging such numbers is meaningless as a matter of statistics. For the average to be meaningful, the difference between "1" and "2" would need to mean the same thing as the difference between "4" and "5." A "1" would have to balance a "5" to be the equivalent of two "3"s. But adding or subtracting labels from each other does not make sense, any more than it makes sense to add or average postal codes.

30. Reporting the averages to several significant digits gives the illusion of scientific precision when in fact the result is not valid.

31. Even if the numbers denoted quantities rather than numerical labels, reducing the data to averages obscures variation, which is crucial for interpreting the information. Even if the ratings had a sensible quantitative interpretation, a class in which student ratings are equally divided between "poor" and "excellent" is presumably quite different from one in which students unanimously rate the instructor "average."

32. One way to reduce the temptation to average labels is to avoid using numerical labels the first place, and instead to tally the frequency of each response category. Of course, if the survey item does not measure what it purports to measure, the frequency distribution of responses will still be misleading.

33. In interpreting survey results, it is important to consider the response rate—the percentage of students in the class who return the survey. Responders are not a random sample: they are "self-selected." When response rates are low, responses are unlikely to be representative of the class as a whole. Responders and nonresponders generally differ. Because anger is generally a stronger motivator than contentment, it is plausible that survey responses are strongly biased towards negative results when response rates are especially low. In short, there is no basis for extrapolating SET or student comments from the students who responded to other students in a course.

34. Response rates themselves are not a mark of teaching quality (Stark and Freishtat, 2014).

35. Student comments need to be interpreted cautiously. Students often use adjectives differently from how faculty do. This includes adjectives such as “fair,” “professional,” “organized,” “challenging,” and “respectful” (Lauer, 2012).

36. Students are generally unable to assess of the appropriateness of material, the organization of the material, the value of the material, or what they have learned (see, e.g., Stark and Freishtat, 2014, and references therein).

37. In my opinion, items relating to teaching effectiveness, course effectiveness, course organization, course relevance, and so on should be eliminated from SET, because these are particularly susceptible to bias and, evidently, misleading. Only items that report students’ experience should be retained, for instance, whether the student enjoyed the class, whether the student found the instructor’s handwriting legible, whether the student found the class easy or difficult, whether the workload was greater than or less than that of other courses, and whether the student has greater or less interest in the subject after taking the class. The results still need to be interpreted and used cautiously. The results should not be reduced to averages. Instead, frequency distributions should be reported: the percentage of students whose response is in each category. Response rates should be reported. Results should not be extrapolated from responders to nonresponders. Results should not be compared across course formats, levels, topics, or disciplines. And the use of the results in employment decisions should be discouraged, if not forbidden: even for such items, responses are likely to be affected by all the confounding biases discussed above, so reliance on SET is likely to have disparate impact on protected groups and to disadvantage some instructors for reasons beyond their control.

III. REFERENCES

30. For supporting references, please see the following, and citations therein:

Ambady, N., and R. Rosenthal, 1993. Half a Minute: Predicting Teacher Evaluations from Thin Slices of Nonverbal Behavior and Physical Attractiveness, *J. Personality and Social Psychology, 64,* 431-441.

Arbuckle, J. and B.D. Williams, 2003. Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations, *Sex Roles, 49,* 507-516. DOI 10.1023/A:1025832707002

Archibeque, O., 2014. Bias in Student Evaluations of Minority Faculty: A Selected Bibliography of Recent Publications, 2005 to Present. http://library.auraria.edu/content/bias-student-evaluations-minority-faculty (last retrieved 30 September 2016)

Basow, S., S. Codos, and J. Martin, 2013. The Effects of Professors' Race and Gender on Student Evaluations and Performance, *College Student Journal, 47* (2), 352-363.

Bedard, K., and P. Kuhn, 2005. Where Class Size Really Matters: Class Size and Student Ratings of Instructor Effectiveness, Department of Economics, University of California, Santa Barbara. http://econ.ucsb.edu/~kelly/ucsb4.pdf (last retrieved 6 October 2016)

Bianchini, S., F. Lissoni, and M. Pezzoni, 2013. Instructor Characteristics and Students' Evaluation of Teaching Effectiveness: Evidence from an Italian Engineering School. *European Journal of Engineering Education, 38,* 38-57. DOI: 0.1080/03043797.2012.742868

Boring, A., 2015. Gender Bias in Student Evaluations of Teachers, OFCE-PRESAGE-Sciences-Po Working Paper, http://www.ofce.sciences-po.fr/pdf/dtravail/WP2015-13.pdf (last retrieved 30 September 2016)

Boring, A., K. Ottoboni, and P.B. Stark, 2016. Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness, *ScienceOpen*, DOI 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Braga, M., M. Paccagnella, and M. Pellizzari, 2014. Evaluating Students' Evaluations of Professors, *Economics of Education Review, 41,* 71-88.

Carrell, S.E., and J.E. West, 2010. Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors, *J. Political Economy, 118,* 409-432.

Hill, M.C., and K.K. Epps, 2010. The Impact of Physical Classroom Environment on Student Satisfaction and Student Evaluation of Teaching in the University Environment, *Academy of Educational Leadership Journal, 14*, 65-

Johnson, V.E., 2003. *Grade Inflation: A Crisis in College Education*, Springer-Verlag, NY, 262pp.

Lake, D.A., 2001. Student Performance and Perceptions of a Lecture-based Course Compared with the Same Course Utilizing Group Discussion. *Physical Therapy, 81*, 896-902.

Lauer, C., 2012. A Comparison of Faculty and Student Perspectives on Course Evaluation Terminology, in *To Improve the Academy: Resources for Faculty, Instructional, and Educational Development, 31,* J.E. Groccia and L. Cruz, eds., Jossey-Bass, 195-211.

MacNell, L., A. Driscoll, and A.N. Hunt, 2015. What's in a Name: Exposing Gender Bias in Student Ratings of Teaching, *Innovative Higher Education, 40*, 291-303. DOI 10.1007/s10755-014-9313-4

Marsh, H.W., and T. Cooper. 1980. Prior Subject Interest, Students Evaluations, and Instructional Effectiveness. Paper presented at the annual meeting of the American Educational Research Association.

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

Schmidt, B., 2015. Gendered Language in Teacher Reviews, http://benschmidt.org/profGender (last retrieved 30 September 2016)

Stark, P.B., and R. Freishtat, 2014. An Evaluation of Course Evaluations, *ScienceOpen*, DOI 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Subtirelu, N.C., 2015. "She does have an accent but…": Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com, *Language in Society, 44,* 35-62. DOI 10.1017/S0047404514000736

Uttl, B., C.A. White, and A. Morin, 2013. The Numbers Tell it All: Students Don't Like Numbers!, *PLoS ONE, 8*(12): e83443, DOI 10.1371/journal.pone.0083443

Uttl, B., C.A. White, and D.W. Gonzalez, 2016. Meta-analysis of Faculty's Teaching Effectiveness: Student Evaluation of Teaching Ratings and Student Learning Are Not Related, *Studies in Educational Evaluation,* DOI: 0.1016/j.stueduc.2016.08.007

Vasta, R., and R.F. Sarmiento, 1979. Liberal Grading Improves Evaluations but not Performance, *J. Educational Psychology, 71*, 207-211.

Williams, W.M., and Ceci, S.J., 1997. "How'm I doing?": Problems with Student Ratings of Instructors and Courses, *Change: The Magazine of Higher Learning*, *29*, 12-23. DOI: 10.1080/00091389709602331

Wolbring, T., and P. Riordan, 2016. How Beauty Works. Theoretical Mechanisms and Two Empirical Applications on Students' Evaluations of Teaching, *Social Science Research, 57*, 253-272. DOI: 10.1016/j.ssresearch.2015.12.009

Worthington, A.C., 2002. The Impact of Student Perceptions and Characteristics on Teaching Evaluations: A Case Study in Finance Education. *Assessment and Evaluation in Higher Education, 27*, 49-64.

10 October 2016

Philip B. Stark