The Effects of Instructor Gender on Student Ratings of Instruction Across Academic Disciplines

Dan Li and Stephen L. Benton

The IDEA Center

Jason Barr

Salus University

Paper presented at the Annual Meeting of the American Educational Research Association, San Antonio, April 2017. Correspondence should be directed to Stephen Benton, Senior Research Officer, The IDEA Center, 301 South Fourth St., Suite 200, Manhattan, KS 66506.

Abstract

This study investigated whether instructor gender interacts with academic discipline group on student ratings of instruction in post-secondary classes. Specifically, we examined gender differences between discipline groups representing science, technology, engineering, and mathematics (STEM) and non-STEM fields, as well across STEM fields. Aggregated IDEA Student Ratings of Instruction (SRI) data from 18,424 female and 15,651 male college and university instructors were analyzed using multivariate analysis of covariance. The statistical analyses revealed extremely weak and practically negligible interaction effects of Instructor Gender × Academic-Discipline Group, $\eta^2_p = .002$, and Instructor Gender × STEM Field, $\eta^2_p = .005$. Although the extent to which student ratings differed by instructor gender depended weakly on the academic group and the STEM field instructors taught in, the gender differences were so trivial that they should not affect decisions made about teaching effectiveness.

The Effects of Instructor Gender on Student Ratings of Instruction Across Academic Disciplines

Although women constitute nearly half of faculty members in the United States (National Center for Education Statistics, 2016), their presence in certain STEM (science, technology, engineering, and mathematics) fields is still low (Committee on Equal Opportunities in Science and Engineering, 2015). Moreover, women tend to be at a disadvantage in comparison with their male peers in terms of academic ranks and income (National Center for Education Statistics, 2009; Snyder, de Brey, & Dillow, 2016). Given its substantial weight in personnel decisions regarding pay raises, contract renewal, and promotion and tenure, an objective and impartial faculty evaluation procedure is critical for closing the workforce gender gap. In this study, we examined the extent to which instructor gender interacts with academic discipline group on student ratings of instruction (SRI), an integral element in faculty evaluation for teaching effectiveness. We are particularly interested in whether a pattern of gender differences, if it exists, is consistent across the still male-dominated STEM fields and their non-STEM counterparts. We also examined whether student ratings vary by instructor gender across the four STEM fields.

Literature Review

Dozens of studies have examined, through experiments and observations, the extent to which student perceptions of teaching effectiveness are related to instructor gender and its interaction with other student, instructor, and course characteristics. Existing experimental studies have compared student ratings of fictitious instructors whose teaching behaviors and characteristics were presented or manipulated through teaching materials and audio or video recordings (K. J. Anderson, 2010; Arbuckle & Williams, 2003; Basow, Codos, & Martin, 2013;

Dukes & Victoria, 1989; Fandt & Stevens, 1991; Freeman, 1994; Haemmerlie & Highfill, 1991; Kaschak, 1978; Kierstead, D'Agostino, & Dill, 1988). Whereas the majority of early experimental research reported no effects from instructor gender on students' perceptions (Feldman, 1992), conflicting findings have since emerged, partially as a result of differences in methods and idiosyncrasies of study samples. For example, MacNell, Driscoll, and Hunt (2014) compared student ratings of four online course sections, where grading and interaction with students were performed by either of two assistant instructors of different gender. For one section each, the instructors falsely identified their gender, thereby creating "actual gender" and "reported gender" course sections. Although no significant differences in student ratings were found between classes taught by the actual female and male teachers, the ratings of the reported male were significantly higher than those of the reported female. However, the experimental design did not establish the equivalence of teaching behaviors performed by the instructors under their real and disguised gender identities, and researcher expectancy effects were not controlled. The conclusion that higher ratings received by perceived male instructors should be attributed to gender bias in student ratings is, therefore, not warranted (see Benton & Li, 2014 for a critique). It is also worth noting that among the few studies that reported effect sizes of significant findings, the variance in student ratings explained by instructor gender or its interaction with student gender was usually too small to indicate practical significance (K. J. Anderson, 2010; Basow et al., 2013; Basow & Silberg, 1987; Dukes & Victoria, 1989; Haemmerlie & Highfill, 1991).

In contrast to experimental studies, observational studies are based on actual student ratings data in classroom settings, which provide larger samples and generally lend more statistical power to detect gender differences. Consequently, observational research tends to

report statistically significant, albeit practically negligible instructor-gender effects on student perceptions of teaching effectiveness (Feldman, 1993). Smith and colleagues (Smith, Yoo, Farr, Salmon, & Miller, 2007), for example, analyzed ratings from approximately 12,000 students in three communication departments, and reported gender effects with trivial effect sizes. Although female instructors were slightly favored by student ratings, the researchers attributed the statistical significance mainly to the large sample size, and suggested administrators should assume gender equality in teaching abilities between male and female instructors during evaluation. Similar preferences for female instructors as reflected in student ratings were reported in studies of business (Caines & Shurden, 2001) and introductory economics courses (K. H. Anderson & Siegfried, 1999). In contrast, using data from over 400 instructors at a U.S. southwest institution, Sidanius and Crane (1989) reported lower ratings received by female instructors on global teaching effectiveness and competency, but they also cautioned that the differences were too small to influence job evaluations. In addition to studies that reported significant results, two large-scale studies in the fields of engineering (Johnson, Narayanan, & Sawaya, 2013) and business (Miles & House, 2015) reported an absence of systemic instructor-gender differences in student ratings.

Several studies examined the effects of instructor gender on SRI by taking academic discipline into account. In a study of 136 instructors at a liberal arts college, Basow (1995) reported a statistically significant three-way interaction effect of divisional affiliation (i.e., academic discipline), instructor gender, and student gender: Ratings were lower to female instructors, by male students, and for natural-science instructors, with the relationships further qualified by the interactions of the three factors. The main effect of discipline was the strongest, although the effect sizes and mean differences were generally small. In a replication study with a

sample of 43 instructors, Basow and Montgomery (2005) reported similar patterns of effects. However, the previously significant Teacher-Gender × Student-Gender interaction effect was no longer present. Centra and Gaubatz (2000) conducted the most comprehensive examination of interaction effects between instructor gender and discipline on SRI. They examined instructor gender differences in courses taught across eight discipline groups: health sciences, business, education, social sciences, fine arts, natural sciences, technology, and humanities. Student ratings for male and female instructors did not differ significantly within any of the discipline categories.

Purpose of the Study

The purpose of this study was to investigate whether instructor gender interacts with academic discipline group on student ratings of instruction in post-secondary classes. We examined whether gender differences would exist in SRI for instructors teaching in STEM and non-STEM fields, given the disproportionate representation of female instructors in the latter group (Hill, Corbett, & St Rose, 2010) and the relatively more challenging nature of some STEM courses (Hativa, 2014). Previous research has confirmed the important roles students' motivation to take the course and their general work habits play in affecting SRI (Benton, Li, Brown, Guo, & Sullivan, 2015). Moreover, students also tend to perceive science and mathematics courses as more difficult, and they express less motivation to take them (Hoyt & Lee, 2002). Consequently, we employed two student characteristics—work habits and motivation to take the course—as covariates in the current study. Specifically, we asked the following research questions:

RQ1. With students' course motivation and work habits controlled, do student ratings on overall measures of teaching effectiveness (i.e., progress on relevant learning objectives, the

overall excellence of the teacher and course) differ by instructor gender in STEM and non-STEM discipline groups?

RQ2. With students' course motivation and work habits controlled, do student ratings on overall measures of teaching effectiveness differ by instructor gender across STEM fields (Science, Technology, Engineering, and Mathematics)?

Method

Variables of interest were collected through the IDEA Student Ratings of Instruction, a two-form survey system that obtains inputs from instructors and students, respectively. Using the *Faculty Information Form* (FIF), instructors indicate the extent to which they emphasized 12 learning objectives in their class by rating each objective as "Minor or No Importance, " "Important, " or "Essential." On the student forms (i.e., *Diagnostic Form* and *Short Form*), students report the progress they made during the course on the same 12 learning objectives, using a scale ranging from 1 (*No apparent progress*) to 5 (*Exceptional progress: I made outstanding gains on this objective*). Additionally, students respond to items measuring the observed frequency of teaching methods practiced by their instructor, as well as related student and course characteristics. Among those are the two covariates that have been found to be related to the three summary measures on IDEA SRI: course motivation ("I really wanted to take this course regardless of who taught it") and work habits ("As a rule, I put forth more effort than other students on academic work.") Both items are measured by a 5-point scale (1 = *Definitely False* and 5 = *Definitely True*).

Student ratings of teaching effectiveness are operationalized by the three overall summary measures on IDEA SRI. Progress on Relevant Objectives (PRO), a weighted mean of average student ratings on relevant learning objectives, is calculated by double weighting student

progress on instructor-identified "essential" objectives and single weighting progress on "important" objectives. The other two summary measures are "Overall, I rate this instructor an excellent teacher," and "Overall, I rate this course as excellent." The scale ranges from 1 (*Definitely False*) to 5 (*Definitely True*).

One of the two independent variables in this study was instructor gender. Since IDEA SRI does not include demographic questions, we inferred instructor gender from their first names, which were collected through the FIF administered online[1]. Personal names are considered as a generally reliable indicator of gender and have been used to predict gender when such information was not readily available (see Lariviere, Ni, Gingras, Cronin, & Sugimoto, 2013 for a bibliometric study on gender inequality in scientific publications). We predicted instructor gender using an R package "gender" (Version 0.5.1; Mullen, 2015), which analyzes historical demographic data to calculate the gender proportion of individuals with a given name and a range of birth year (Blevins & Mullen, 2015). We chose the historical dataset (1930 to 2012) from the U.S. Social Security Administration and specified the range of birth years from 1932 to 1990, which were conservative estimates given the surveys were administered in 2002 to 2015. To mitigate ambiguities introduced by gender-neutral names, we retained only courses where the predicted proportion of one gender was at least 90%, and assigned the predominant gender as the prediction.

The other independent variable was the discipline group of non-STEM and STEM fields, with the latter further broken down into science, technology, engineering, and mathematics respectively. The FIF asks instructors to indicate the discipline code for their course, which is a

---

[1] The paper version of FIF collects only instructors' last name (up to 11 letters) and initials due to space restriction.

four-digit record similar to The Classification of Instructional Programs (CIP) created by the National Center for Education Statistics. We first recoded the discipline code variable to match the 47 broad disciplines defined in CIP and then grouped relevant disciplines into the following fields: science (agriculture, physical sciences, and biological sciences), technology (computer and information sciences), engineering (engineering, and engineering technologies and engineering-related fields), mathematics (mathematics and statistics), as well as non-STEM (all other disciplines).

Data source

We chose the annual IDEA SRI research datasets as the data source. For the purpose of this study, we selected course-level data collected online from 2002 to 2015. The datasets are comprised of student ratings aggregated at the course level, collected through paper-and-pencil and online surveys. Included in the research datasets are course data collected from more than 300 institutions, which are geographically spread across the U.S. and fall into all major categories of Carnegie Classifications (i.e., Associates, Baccalaureate, Master's, and Doctoral). Several exclusions were performed to create a research dataset that maximized score reliability and representativeness of the population. Courses with fewer than 10 responses were excluded due to low reliability. Classes where the instructor failed to identify any relevant learning objectives were also removed. In addition, random courses were dropped to ensure no institution constituted more than 5% to the entire sample. Since about 60% of instructors in the analytic sample had multiple course records, we defined the unit of analysis as the average student ratings an instructor received across courses in the same discipline group (i.e., science, technology, engineering, mathematics, or non-STEM).

To compute average ratings each instructor received, we used a combination of the instructor's first name, surname, and a unique identifier of the institution as an instructor ID for each course record. Courses sharing the same ID were averaged to create the mean ratings at the instructor level. Courses with a response rate lower than 50% were removed from the analytic sample to reduce bias from less representative course samples. We also excluded instructors who had course records in multiple discipline groups to ensure the independence of observations. As a result, ratings of 34,075 instructors were included in the analytic sample.

Sample Description

The analytic sample included 34,075 instructors from 280 U.S. institutions. Eight-four percent of instructors taught in non-STEM fields (n = 28,631), where female instructors outnumbered male instructors (57% vs. 43%). Among the 5,444 STEM instructors, slightly more than one third (37%) were women. While the gender gaps in mathematics and science were relatively small (46% and 40% females respectively), men were approximately two and five times the proportion of women respectively in technology and engineering. Tables 1 and 2 display the proportions of male and female instructors categorized by discipline groups and STEM fields respectively.

Data Analysis

We first conducted a 2 × 2 (Gender [male, female] × Discipline Group [STEM, non-STEM]) between-subjects multivariate analysis of covariance (MANCOVA) to examine differences on the three overall summary measures, controlling for the influence of students' course motivation and work habits. To investigate gender differences within STEM fields, we performed a 2 × 4 (Gender × STEM Field [science, technology, engineering, mathematics]) MANCOVA with the same two covariates. For all analyses, we reported partial eta squared (η^2_p)

as a measure of effect size, which denotes proportion of variance accounted for in dependent variables. Pillai's trace was selected as the multivariate criterion due to its robustness to the unequal sample sizes in this study. Univariate analyses were conducted following any significant multivariate effect, and the Tukey test was applied for post-hoc comparisons. An important assumption of analysis of covariance is the homogeneity of covariate regression slopes. To test for this assumption, we compared the slopes for each of the three overall measures regressed on each covariate between male and female instructors, discipline group, and STEM fields. There was no evidence that the slopes varied meaningfully by the groups.

Results

RQ1 examines whether student ratings on progress on relevant learning objectives, excellence of the teacher, and excellence of the course differ by instructor gender in STEM and non-STEM discipline groups, with the effects of course motivation and work habits controlled. Descriptive statistics for the dependent variables and covariates grouped by instructor gender and discipline group are presented in Table 3.

The results of the MANCOVA and subsequent univariate analyses are summarized in Table 4. Alpha was set at .01 given the large sample size. Both covariates were significantly related to the dependent variables with considerable effect sizes, $F(3, 34072) = 6{,}207.91$, $p < .001$, $\eta^2_p = .35$ for course motivation and $F(3, 34072) = 1{,}702.54$, $p < .001$, $\eta^2_p = .13$ for work habits, confirming the need to control for their influence. Table 5 demonstrates adjusted mean scores and standard errors of the overall summary measures.

The multivariate interaction effect between instructor gender and discipline group on student ratings was statistically significant, $F(3, 34072) = 19.49$, $p < .001$, although the effect size was trivial ($\eta^2_p = .002$). The univariate analyses demonstrated that the interaction effect

resided weakly in PRO ($\eta^2_p < .001$), excellence of the teacher ($\eta^2_p = .001$), and excellence of the course ($\eta^2_p = .002$). That is, the interaction effect accounted for no more than 0.2% of the variance in each dependent variable. As shown in Table 5, the main effect of gender on student ratings varies depending on the discipline group. After adjusting for the effects of course motivation and work habits, male non-STEM instructors tended to receive slightly higher ratings than their female colleagues, while female STEM teachers were rated slightly better than their male peers. The gender differences in the adjusted means of overall summary measures, ranging from 0.02 to 0.06, were statistically significant but trivial. Although the multivariate main effect of instructor gender was statistically significant, $F(3, 34072) = 15.83$, $p < .001$, the effect size was too small to indicate practical significance ($\eta^2_p = .001$). Moreover, only the univariate test on one of the three dependent measures, namely, progress on relevant objectives, reached the .01 level of significance with a negligible effect size, $F(1, 34074) = 7.05$, $p = .008$, $\eta^2_p < .001$.

RQ2 extends RQ1 by focusing on whether gender differences exist across the four fields of STEM, after adjusting for the influence of course motivation and work habits. Descriptive statistics of variables are shown in Table 6. Table 7 displays the results of the MANCOVA and subsequent univariate analyses. After controlling for the effects of the covariates, adjusted means and standard errors are presented in Table 8. Similar to RQ1, the multivariate and univariate interaction effects between instructor gender and STEM fields on student ratings were all statistically significant, with effect sizes ranging from .003 to .005. The extent to which student ratings of male and female instructors varied depends on which field in STEM they taught. Results of independent t-tests with a Tukey correction show that after adjusting for the effects of course motivation and work habits, female instructors in Technology received slightly higher ratings than their male colleagues, with the gender gap being about 0.10 for the three overall

summary measures. Similarly, student ratings for female instructors in Mathematics were higher than those for male instructors, with the greatest difference being 0.10 for excellence of teacher and the smaller gap of 0.06 for the other two measures. In contrast, the gender differences in Science and Engineering were not statistically significant. The main effect of instructor gender in STEM fields mirror the results of RQ1, with statistical significance and trivial effect sizes for the multivariate analysis, $F(3, 34072) = 5.62$, $p < .001$, $\eta^2_p = .003$, and the univariate analysis of PRO, $F(1, 34074) = 9.40$, $p = .002$, $\eta^2_p = .002$.

Discussion

The results of the present study can be summarized as follows. First, as indicated by negligible effect sizes and trivial gender differences, instructor gender had no practically meaningful effects on student ratings of overall summary measures, in the comparison of instructors teaching in STEM and non-STEM fields. Students rated their overall progress, the quality of the teacher, and the excellence of the course very similarly regardless of whether they were taught by a man or a woman. Second, across the STEM fields, students assigned slightly higher ratings of the three overall measures for female instructors in Technology and Mathematics, while no gender differences were found for SRI of Science and Engineering instructors. Third, course motivation and work habits were important covariates that should be taken into account when measuring learning outcomes.

The lack of meaningful differences in the ratings of female and male instructors on overall summary measures in STEM and non-STEM groups supports the results of previous large-scale research (Centra, 2009; Centra & Gaubatz, 2000; Feldman, 1993). The exceptionally weak interaction effects found between instructor gender and discipline group suggest that in the natural settings, neither gender is superior in teaching, in either STEM or non-STEM fields. The

findings that women in Technology and Mathematics tended to receive higher student ratings echo existing research in the fields of communication (Smith et al., 2007), business (Caines & Shurden, 2001), and economics (K. H. Anderson & Siegfried, 1999). However, since the greatest average gender difference was 0.11 on a five-point scale and student ratings is only one of multiple sources of evidence for teaching evaluation, more empirical support is needed to claim female instructors in the two fields are better teachers.

Results of the study also suggest that gender bias in student ratings as found in previous research (for a critique, see Benton & Li, 2014; e.g., Boring, 2017; Boring, Ottoboni, & Stark, 2016; MacNell et al., 2014; and Ryalls, Benton, Barr, & Li, 2015) may be more an artifact of research design than students' favoritism of one gender over the other. When gender differences have been found in SRI, they have usually occurred in laboratory studies, where students rated descriptions of fictitious teachers who varied in gender (Feldman, 1992). In contrast, in studies conducted on ratings of actual teachers in the classroom, researchers have found, as we did, no meaningful differences due to gender or only a very weak relationship that favors female instructors (Bennett, 1982; Centra, 2009; Feldman, 1993; Smith et al., 2007). As Feldman (1992) concludes, "Any predispositions of students in the social laboratory to view male and female college teachers in certain ways (or the lack of such predispositions) may be modified by students' actual experiences with their teachers in the classroom or lecture hall" (p. 152). Feldman's assertion is consistent with Gordon Allport's (1954) *contact theory,* which posits that actual personal interaction can override stereotypes and reduce biases, a view supported more recently by others (Amichai-Hamburger & McKenna, 2006; Pettigrew & Tropp, 2006).

The analyses of student ratings collected through IDEA SRI indicate that a properly designed evaluation instrument can mitigate biases that threaten the validity of the measurement.

Previous research suggests that student expectations of gender roles may account for gender bias in SRI (Andersen & Miller, 1997; Bachen, McLoughlin, & Garcia, 1999). Items worded in a neutral manner that do not embrace gender-stereotyped teaching practices may effectively reduce potential bias associated with student expectations.

The absence of meaningful gender differences in this study does not necessarily mean that gender bias does not exist in the practice of faculty evaluation. Faculty evaluation is a holistic procedure that involves multiple sources of evidence obtained through various channels. Therefore, at various stages of the process faculty evaluation and consequent personnel decisions are prone to biases, which are inherent in individual and collective perceptions and expectations of certain demographic or cultural groups. Without fair means of collecting and using evaluation evidences, gender bias may well systematically harm one gender through individual student ratings, peer evaluation from other faculty members, the decisions of administrators, and inputs from other parties. However, this study discovers no favoritism toward either gender in aggregated student ratings that is strong enough to systematically influence teacher evaluations, *as long as student ratings do not serve as the only measure of teaching effectiveness and administrators do not make too much of too little*.

Another important finding from the current study is that the strong and positive effects of course motivation and work habits exhibited as covariates on SRI validate the need to control for circumstances that may affect student ratings but are beyond the instructor's control. Certain disciplines are inherently more challenging and require more devotion from students and instructors, which should also be taken into consideration when SRI are reviewed.

Limitations

This study is not without limitations. First, the research data set was not based on a randomly selected sample, and thus findings may not be applicable to all disciplines and institutions. Nonetheless, the sample was large and included courses from all major Carnegie classifications and from numerous disciplines and institutions. Second, whereas inferring gender based on first names has become an increasingly common practice when direct measures of gender are absent, its drawbacks should be taken into account. Instructors with gender-neutral or uncommon names, as well as those from cultures where first names are less gender-typed, may be underrepresented in the sample due to uncertainty in estimation. Third, although student gender has been suggested by previous research as a covariate for SRI, this study did not control for it because the student forms were anonymous.

Implications

The effects of instructor gender and its interaction with academic-discipline group do not exert much influence on overall IDEA SRI measures. The greatest differences in ratings are observed between men and women teaching technology and mathematics, although the gender gap was too small to claim either gender is comprised of better teachers. When properly used as one of multiple sources of evidence, mean class scores on IDEA SRI are a meaningful measure of student perceptions of teaching effectiveness and suggest more gender equality than differences in teaching quality and behaviors. Nonetheless, IDEA users may want to examine this issue on their own campuses. At local levels, some differences could be meaningful, particularly if ratings are used exclusively in making summative decisions about teaching effectiveness.

Despite the statistical significance of gender effects due to the large sample size, our analysis found no practical gender differences in summary measures of SRI. Student ratings of their progress on relevant objectives, excellence of teacher, and excellence of course were remarkably similar between female and male instructors when compared within the same discipline group. It is particularly encouraging that PRO, an exceptionally valid and reliable measure of teaching effectiveness, confirms the gender equality in teaching abilities. When properly used as one of the multiple sources of evidence, we do not expect SRI to be a source of gender bias in faculty evaluation.

References

Allport, G. W. (1954). The nature of prejudice. Cambridge, MA: Perseus Books.

Amichai-Hamburger, Y., & McKenna, K. Y. A. (2006). The contact hypothesis reconsidered: Interacting via the Internet. *Journal of Computer-Mediated Communication*, *11*(3), 825–843. http://doi.org/10.1111/j.1083-6101.2006.00037.x

Andersen, K., & Miller, E. D. (1997). Gender and student evaluations of teaching. *Political Science & Politics*, *30*(2), 216–219. http://doi.org/10.2307/420499

Anderson, K. H., & Siegfried, J. J. (1999). Gender differences in rating the teaching of economics. *Eastern Economic Journal*, *23*(3), 347–473.

Anderson, K. J. (2010). Students' stereotypes of professors: An exploration of the double violations of ethnicity and gender. *Social Psychology of Education*, *13*(4), 459–472. http://doi.org/10.1007/s11218-010-9121-3

Arbuckle, J., & Williams, B. D. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles*, *49*(9), 507–516. http://doi.org/10.1023/A:1025832707002

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education*, *48*(3), 193–210. http://doi.org/10.1080/03634529909379169

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology*, *87*(4), 656–665. http://doi.org/10.1037/0022-0663.87.4.656

Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education*, *18*(2), 91–106. http://doi.org/10.1007/s11092-006-9001-8

Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology*, *79*(3), 308–314. http://doi.org/10.1037/0022-0663.79.3.308

Basow, S. A., Codos, S., & Martin, J. (2013). The effects of professors' race and gender on student evaluations and performance. *College Student Journal*, *47*(2), 352–363.

Bennett, S. K. (1982). 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. http://doi.org/10.1037/0022-0663.74.2.170

Benton, S. L., & Li, D. (2014). What's in the study: Exposing validity threats in the MacNell, Driscoll, and Hunt study of gender bias. Retrieved January 27, 2017, from http://www.ideaedu.org/Resources-Events/IDEA-Blog/PostId/47/whats-in-the-study-exposing-validity-threats-in-the-macnell-driscoll-and-hunt-study-of-gender-bias

Benton, S. L., Li, D., Brown, R., Guo, M., & Sullivan, P. (2015). *IDEA Technical Report No. 18: Revising the IDEA Student Ratings of Instruction system, 2002-2011 Data*. Manhattan, KS: The IDEA Center.

Blevins, C., & Mullen, L. (2015). Jane, John … Leslie? A historical method for algorithmic gender prediction. *Digital Humanities Quarterly*, *9*(3). Retrieved from http://www.digitalhumanities.org/dhq/vol/9/3/000223/000223.html

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, *145*, 27–41. http://doi.org/10.1016/j.jpubeco.2016.11.006

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

Caines, W. R., & Shurden, M. C. (2001). Gender issues in the student ratings of school of business instructors at a regional university. *Academy of Educational Leadership Journal*, *5*(2), 39–46.

Centra, J. A. (2009). *Differences in responses to the Student Instructional Report: Is it bias*. Princeton, NJ: Educational Testing Service.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education*, *71*(1), 17–33. http://doi.org/10.1080/00221546.2000.11780814

Committee on Equal Opportunities in Science and Engineering. (2015). *Committee on Equal Opportunities in Science and Engineering 2013-2014 biennial report to Congress: Broadening participation in America's STEM workforce*. National Science Foundation.

Dukes, R. L., & Victoria, G. (1989). The effects of gender, status, and effective teaching on the evaluation of college instruction. *Teaching Sociology*, *17*(4), 447–457. http://doi.org/10.2307/1318422

Fandt, P. M., & Stevens, G. E. (1991). Evaluation bias in the business classroom: Evidence relating to the effects of previous experiences. *The Journal of Psychology*, *125*(4), 469–477. http://doi.org/10.1080/00223980.1991.10543309

Feldman, K. A. (1992). College students' views of male and female college teachers: Part I—Evidence from the social laboratory and experiments. *Research in Higher Education*, *33*(3), 317–375. http://doi.org/10.1007/BF00992265

Feldman, K. A. (1993). College students“ views of male and female college teachers: Part II: Evidence from students” evaluations of their classroom teachers. *Research in Higher Education*, *34*(2), 151–211.

Freeman, H. R. (1994). Student evaluations of college instructors: Effects of type of course taught, instructor gender and gender role, and student gender. *Journal of Educational*

Psychology, *86*(4), 627–630. http://doi.org/10.1037/0022-0663.86.4.627

Haemmerlie, F. M., & Highfill, L. A. (1991). Bias by male engineering undergraduates in their evaluation of teaching. *Psychological Reports*, *68*(1), 151–160. http://doi.org/10.2466/pr0.1991.68.1.151

Hativa, N. (2014). Student ratings of instruction: Recognizing effective teaching (Second Edition). Oron Publications.

Hill, C., Corbett, C., & St Rose, A. (2010). *Why so few? Women in Science, Technology, Engineering, and Mathematics* (pp. 1–134). American Association of University Women.

Hoyt, D. P., & Lee, E.-J. (2002). *Technical Report No. 13: Disciplinary differences in student ratings*. Manhattan, KS: The IDEA Center. Retrieved from http://www.ideaedu.org/Portals/0/Uploads/Documents/Technical-Reports/Disciplinary-Differences-in-Student-Ratings_techreport-13.pdf

Johnson, M. D., Narayanan, A., & Sawaya, W. J. (2013). Effects of course and instructor characteristics on student evaluation of teaching across a college of engineering. *Journal of Engineering Education*, *102*(2), 289–318. http://doi.org/10.1002/jee.20013

Kaschak, E. (1978). Sex bias in student evaluations of college professors. *Psychology of Women Quarterly*, *2*(3), 235–243. http://doi.org/10.1111/j.1471-6402.1978.tb00505.x

Kierstead, D., D'Agostino, P., & Dill, H. (1988). Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology*, *80*(3), 342–344. http://doi.org/10.1037/0022-0663.80.3.342

Lariviere, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*(7479), 211–213. http://doi.org/10.1038/504211a

MacNell, L., Driscoll, A., & Hunt, A. N. (2014). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, *40*(4), 291–303. http://doi.org/10.1007/s10755-014-9313-4

Miles, P., & House, D. (2015). The tail wagging the dog; An overdue examination of student teaching evaluations. *International Journal of Higher Education*, *4*(2), 1–11. http://doi.org/10.5430/ijhe.v4n2p116

Mullen, L. (2015). gender: Predict gender from names using historical data.

National Center for Education Statistics. (2009, January). Table 315.60. Full-time and part-time faculty and instructional staff in degree-granting postsecondary institutions, by race/ethnicity, sex, and selected characteristics: Fall 2003. National Center for Education Statistics. Retrieved from https://nces.ed.gov/programs/digest/d15/tables/dt15_315.60.asp

National Center for Education Statistics. (2016, March). Table 314.20. Employees in degree-granting postsecondary institutions, by sex, employment status, control and level of

institution, and primary occupation: Selected years, fall 1991 through fall 2013. Retrieved from https://nces.ed.gov/programs/digest/d15/tables/dt15_314.20.asp

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751–783. http://doi.org/10.1037/0022-3514.90.5.751

Ryalls, K., Benton, S. L., Barr, J., & Li, D. (2015). *Response to "Bias against female instructors."* Manhattan, KS: The IDEA Center. Retrieved from http://www.ideaedu.org/Portals/0/Uploads/Documents/Response_to_Bias_Against_Female_Instructors.pdf

Sidanius, J., & Crane, M. (1989). Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology*, *19*(2), 174–197. http://doi.org/10.1111/j.1559-1816.1989.tb00051.x

Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The Influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. *Women's Studies in Communication*, *30*(1), 64–77. http://doi.org/10.1080/07491409.2007.10162505

Snyder, T. D., de Brey, C., & Dillow, S. A. (2016). *Digest of education statistics 2015*. National Center for Education Statistics.

Table 1

*Number and Percentage of Female and Male Instructors by Discipline Group (*N = *34,075)*

	Female		Male	
Discipline group	*n*	%	*n*	%
Non-STEM	16,415	57.3	12,216	42.7
STEM	2,009	36.9	3,435	63.1

Note. Percentage within rows unless otherwise specified.

Table 2

*Number and Percentage of Female and Male Instructors by STEM Field (*N = *5,444)*

	Female		Male	
STEM Field	*n*	%	*n*	%
Science	861	40.0	1,289	60.0
Technology	308	32.4	644	67.6
Engineering	142	17.1	688	82.9
Mathematics	698	46.2	814	53.8

Note. Percentage within rows unless otherwise specified.

Table 3

*Mean Scores and Standard Deviations of Dependent Variables and Covariates as a Function of Instructor Gender and Discipline Group (*N = *34,075)*

	Dependent variables						Covariates			
	PRO		Excellence of teacher		Excellence of course		Course motivation		Work habits	
Group	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Non-STEM										
Female	4.14	0.42	4.22	0.59	4.09	0.55	3.69	0.49	3.96	0.26
Male	4.12	0.42	4.25	0.57	4.10	0.54	3.61	0.49	3.93	0.25
STEM										
Female	3.98	0.46	4.09	0.66	3.87	0.57	3.47	0.48	3.88	0.26
Male	3.96	0.47	4.06	0.65	3.86	0.58	3.53	0.48	3.89	0.26

Note. PRO = Progress on Relevant Objectives.

Table 4

Multivariate and Univariate Analyses of Covariance for Instructor Gender and Discipline Group, With Course Motivation and Work Habits as Covariates

				Univariate								
	Multivariate			PRO			Excellence of teacher			Excellence of course		
Source	F^a	p	η^2_p	F^b	p	η^2_p	F^b	p	η^2_p	F^b	p	η^2_p
Course motivation (covariate)	6207.91	< .001	.353	4878.68	< .001	.125	3677.89	< .001	.097	11855.00	< .001	.258
Work habits (covariate)	1702.54	< .001	.130	3279.01	< .001	.088	584.03	< .001	.017	960.52	< .001	.027
Gender	15.83	< .001	.001	7.05	.008	< .001	0.40	.528	< .001	1.39	.238	< .001
Discipline group	156.75	< .001	.014	236.54	< .001	.007	104.27	< .001	.003	297.56	< .001	.009
Gender × Discipline group	19.49	< .001	.002	33.52	< .001	< .001	49.00	< .001	.001	56.66	< .001	.002

Note. Multivariate F ratios were generated from Pillai's statistic. [a]Multivariate $df = 3, 34072$. [b]Univariate $df = 1, 34074$.

Table 5

Adjusted Mean Scores and Standard Errors of Student Ratings of Overall Summary Measures Based on Course Motivation and Work Habits Grouped by Instructor Gender and Discipline Group

	PRO		Excellence of teacher		Excellence of course	
Group	*M*	*SE*	*M*	*SE*	*M*	*SE*
Non-STEM						
Female	4.11	0.003	4.20	0.004	4.05	0.004
Male	4.13	0.003	4.26	0.005	4.11	0.004
STEM						
Female	4.06	0.008	4.17	0.012	3.98	0.010
Male	4.01	0.006	4.11	0.009	3.94	0.008

Note. PRO = Progress on Relevant Objectives.

Table 6

*Mean Scores and Standard Deviations of Dependent Variables and Covariates as a Function of Instructor Gender and STEM Fields (*N = *5,444)*

	Dependent variables						Covariates			
	PRO		Excellence of teacher		Excellence of course		Course motivation		Work habits	
Group	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Science										
Female	4.01	0.49	4.05	0.69	3.89	0.58	3.53	0.46	3.96	0.24
Male	3.98	0.47	4.07	0.66	3.86	0.59	3.49	0.45	3.96	0.25
Technology										
Female	4.04	0.39	4.17	0.58	4.01	0.54	3.64	0.52	3.83	0.26
Male	3.95	0.45	4.08	0.63	3.94	0.55	3.68	0.47	3.84	0.28
Engineering										
Female	3.94	0.49	3.90	0.71	3.83	0.60	3.66	0.40	3.89	0.20
Male	3.99	0.45	4.04	0.63	3.93	0.55	3.72	0.40	3.91	0.22
Mathematics										
Female	3.93	0.46	4.13	0.63	3.79	0.54	3.27	0.45	3.81	0.26
Male	3.89	0.48	4.05	0.67	3.75	0.57	3.31	0.47	3.83	0.25

Note. PRO = Progress on Relevant Objectives.

Table 7

Multivariate and Univariate Analyses of Covariance for Instructor Gender and STEM Fields, With Course Motivation and Work Habits as Covariates

				Univariate								
	Multivariate			PRO			Excellence of teacher			Excellence of course		
Source	F^a	p	η^2_p	F^b	p	η^2_p	F^b	p	η^2_p	F^b	p	η^2_p
Course motivation (covariate)	884.44	< .001	.328	996.91	< .001	.155	744.64	< .001	.121	1991.42	< .001	.268
Work habits (covariate)	281.89	< .001	.135	640.7	< .001	.105	178.79	< .001	.032	264.79	< .001	.046
Gender	5.62	< .001	.003	9.40	.002	.002	1.27	.260	< .001	3.59	.058	< .001
STEM fields	25.11	< .001	.041	18.58	< .001	.010	51.93	< .001	.028	22.32	< .001	.012
Gender × STEM fields	3.24	< .001	.005	4.75	.003	.003	7.49	< .001	.004	5.00	.002	.003

Note. Multivariate F ratios were generated from Pillai's statistic. [a]Multivariate $df = 3, 5441$. [b]Univariate $df = 1, 5443$.

Table 8

Adjusted Mean Scores and Standard Errors of Student Ratings of Overall Summary Measures Based on Course Motivation and Work Habits Grouped by Instructor Gender and STEM Fields

	PRO		Excellence of teacher		Excellence of course	
Group	*M*	*SE*	*M*	*SE*	*M*	*SE*
Science						
Female	3.96	0.01	4.01	0.02	3.84	0.02
Male	3.95	0.01	4.05	0.02	3.84	0.01
Technology						
Female	4.03	0.02	4.14	0.03	3.95	0.03
Male	3.92	0.02	4.02	0.02	3.85	0.02
Engineering						
Female	3.88	0.03	3.83	0.05	3.73	0.04
Male	3.90	0.02	3.92	0.02	3.79	0.02
Mathematics						
Female	4.06	0.02	4.28	0.02	3.97	0.02
Male	4.00	0.01	4.18	0.02	3.91	0.02

Note. PRO = Progress on Relevant Objectives.