IDEA Research Report #10 • April 2017

The Effects of Instructor Gender and Discipline Group on Student Ratings of Instruction

Dan Li and Stephen L. Benton • The IDEA Center

Despite the nearly equal gender ratio of faculty members in the United States (National Center for Education Statistics, 2016), women are still underrepresented in certain STEM (science, technology, engineering, and mathematics) fields (Committee on Equal Opportunities in Science and Engineering, 2015). Also, women tend to hold lower academic ranks (National Center for Education Statistics, 2009) and earn a noticeably lower salary than their male counterparts (Snyder, de Brey, & Dillow, 2016). An objective and impartial faculty evaluation procedure, given its substantial weight in personnel decisions regarding pay raises, contract renewal, and promotion and tenure, is critical for closing the workforce gender gap. In this study, the authors examine the extent to which instructor gender affects student ratings of instruction (SRI), an integral component in evaluating faculty 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.

The Relationship Between Instructor Gender and SRI

Dozens of studies have investigated the extent to which student perceptions of teaching effectiveness relate 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 and characteristics were presented or manipulated through syllabi (Anderson, 2010); description of the instructor (Freeman, 1994; Haemmerlie & Highfill, 1991; Kaschak, 1978; Kierstead, D'Agostino, & Dill, 1988); teaching scenarios (Dukes & Victoria, 1989); and lectures given by actors (Fandt & Stevens, 1991) or animated characters (Arbuckle & Williams, 2003; Basow, Codos, & Martin, 2013). Whereas the majority of early experimental research reported no differences in students' perception based on instructor gender (Feldman, 1992), conflicting findings have since emerged, partially due to differences in methods and idiosyncrasies of study samples. Among the 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 (Anderson, 2010; Basow et al., 2013; Basow & Silberg, 1987; Dukes & Victoria, 1989; Haemmerlie & Highfill, 1991).

Observational studies based on actual student ratings data in classroom settings generally report practically negligible instructor-gender effects on student perceptions of teaching effectiveness (Feldman, 1993). In their analysis of more than 12,000 student individual ratings from three communication departments, Smith and colleagues reported a statistically significant gender effect with trivial effect sizes and suggested administrators should assume equal teaching abilities between male and female instructors (Smith, Yoo, Farr, Salmon, & Miller, 2007). Sidanius and Crane (1989) found that female instructors received lower ratings on global teaching effectiveness and competency, but they also cautioned that the differences were too small to affect job evaluations. Caines and Shurden (2001) compared student ratings of nearly 700 business courses and found that students gave female instructors slightly higher ratings on teaching effectiveness and the use of certain teaching methods. A lack of systemic instructor-gender differences in student ratings was reported in other large-scale studies on engineering (Johnson, Narayanan, & Sawaya, 2013) and business instructors (Miles & House, 2015).

Several studies examined the effects of instructor gender and academic discipline on SRI. In a study of 136 instructors at a liberal arts college, Basow (1995) reported the statistically significant main and interaction effects of divisional affiliation (i.e., academic discipline), instructor gender, and student gender: Ratings were lower for female teachers, male students, and natural-science teachers, 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 Instructor-gender ×

Discipline interactions 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. In our study, we investigated whether gender differences would exist in STEM and non-STEM fields, given the disproportionate representation of female instructors in the latter (Hill, Corbett, & St Rose, 2010) and the relatively more challenging nature of some STEM courses (Hativa, 2014).

The Relationship Between SRI and Academic Discipline

On average, students rate courses in the humanities and arts more highly than those in the social sciences, which in turn are rated more highly than math and science courses (Braskamp & Ory, 1994; Cashin, 1990; Centra, 1993; 2009; Feldman, 1978; Hoyt & Lee, 2002; Kember & Leung, 2011; Marsh & Dunkin, 1997; Sixbury & Cashin, 1995). Authors have offered several explanations for these differences. Courses in some fields may receive lower ratings because they are not as well taught (Cashin, 1990). Students enrolled in mathematics and science courses, for example, report less frequent instructor use of several teaching methods strongly related to global ratings of teaching excellence: stimulating student interest, fostering collaboration, and encouraging student involvement (Benton, Gross, & Brown, 2012). Instructors in soft disciplines, on the other hand, tend to exhibit a wider range of teaching behaviors and foster active learning more than those in hard disciplines (Franklin & Theall, 1992; Hativa, 2014).

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. In addition, courses in fields requiring more quantitative reasoning may receive lower ratings because contemporary students are less competent in such skills. If that is the case, then some control is necessary, as is done with IDEA's discipline comparative scores (Cashin, 1990).

Another explanation concerns the differential structure of content across academic disciplines (Hativa, 2014). Hard disciplines (e.g., engineering, chemistry) are characterized by a structured knowledge sequence organized around a theory accepted by all members of the field (see Biglan, 1973). To succeed in such courses, students must have a solid knowledge base in the content area. Students who perceive their background preparation as inadequate are actually more likely to assign low ratings (Benton, Li, Brown, Guo, & Sullivan, 2015). Finally, some faculty may be attracted to certain disciplines because they offer greater opportunity for research than teaching (Hativa, 2014), and thus those instructors do not prioritize teaching skills.

Relationships Between Overall Measures and Teaching Methods

As close observers in multiple course sessions, students can report their perceptions of the frequency of specific teaching behaviors, termed *teaching methods* in IDEA SRI. Students' ratings of teaching methods are significantly related to students' overall ratings of the teacher and the course as well as to average student progress on relevant course objectives (Benton et al., 2015). In particular, seven of IDEA's teaching methods are highly correlated with ratings of teacher and course excellence and represent four major areas of effective teaching (see Table 1; Benton et al., 2015). Therefore, these seven methods are included in IDEA's *Teaching Essentials* (TE; http://www.ideaedu.org/Services/Teaching-Essentials). In this study, we thus investigated whether students' ratings of how frequently they observed the seven TE methods vary by instructor gender and discipline group.

Purpose of the Study

The purpose of this study was to investigate whether student ratings of male and female instructors differed in IDEA SRI and, if so, whether the differences varied between STEM and non-STEM fields. Specifically, we asked the following research questions.

Research Question 1. After controlling for students' course motivation and work habits, do student ratings on overall measures of teaching effectiveness (i.e., progress on relevant learning objectives and the overall excellence of the teacher and course) differ by instructor gender and discipline group (STEM vs. non-STEM)?

Research Question 2. Do student ratings of teaching methods vary by instructor gender and discipline group (STEM vs. non-STEM)?

Method

Data Source

Data were collected through the IDEA Legacy SRI online platform (http://www.ideaedu.org/Resources-Events/Support-For-Current-Clients/IDEA-Legacy-Online-and-Paper-Platform) from 2002 to 2015, including instructor responses on the Faculty Information Form

Table 1
Teaching Essential *Teaching Methods Related to Overall Summary Measures*

Teaching method category	Overall summary measure	
	Excellence of instructor	Excellence of course
Organization		– Made it clear how each topic fit into the course
Clarity	– Explained course material clearly	– Explained course material clearly
Enthusiasm/expression	– Introduced stimulating ideas about the subject	– Introduced stimulating ideas about the subject – Inspired students to set and achieve goals which really challenged them – Demonstrated the importance and significance of the subject matter
Rapport/interactions	– Displayed a personal interest in students and their learning – Found ways to help students answer their own questions	

(FIF) and course-level mean scores of student responses to items on the Diagnostic Form. Because the majority of instructors 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. To reduce bias introduced by courses with low response rates, we restricted the analytic sample to courses with a response rate of at least 50%. We also excluded instructors who had taught courses in both discipline groups to ensure the independence of observations. As a result, ratings of 25,243 instructors were included in the analytic sample.

Measures

The dependent variables measuring teaching effectiveness are operationalized by three overall summary measures on IDEA SRI. Progress on Relevant Objectives (PRO) is a weighted mean of average student ratings on instructor-identified relevant learning objectives. Using the FIF, instructors indicate the relevance of each of the 12 learning objectives as “minor or no importance,” “important,” or “essential.” PRO is calculated by double weighting course-level average student progress on essential objectives and single weighting progress on important objectives. The other two summary measures are “Overall, I rate this instructor an excellent teacher” (excellence of teacher) and “Overall, I rate this course as excellent” (excellence of course). The scale ranges from 1 (*definitely false*) to 5 (*definitely true*). Using the same scale, mean scores on two student characteristics served as covariates: “I really wanted to take this course regardless of who taught it” (motivation) and “As a rule, I put forth more effort than other students on academic work” (work habits). Student ratings on the frequency of seven teaching methods were collected on a 5-point scale (1 = *hardly ever* and 5 = *almost always*).

Because the FIF does not include demographic questions, we inferred instructor gender, an independent variable, from instructors’ first names, which were collected through the FIFs administered online.[1] We predicted instructor gender using an R package “gender” (Mullen, 2015), which analyzes historical data to calculate the gender proportion of individuals with a given name and a birth year within a given range (Blevins & Mullen, 2015). To mitigate ambiguities introduced by gender-neutral names, we retained only courses where the predicted proportion of one gender was at least 90%. We then assigned the predominant gender as the prediction.

The other independent variable was discipline group. On the FIF, instructors indicate the course’s academic discipline, using a four-digit record similar to the Classification of Instructional Programs (CIP) created by the National Center for Education Statistics. For the

[1] FIFs are administered as Web-based or paper-and-pencil questionnaires. Data collected through the paper version contain only the instructor’s last name (up to 11 letters) and initials, making it impossible to infer instructor gender from first names.

purpose of this study, we grouped courses into STEM and non-STEM.[2]

Sample Description

The analytic sample included 25,243 instructors from 256 U.S. institutions. Among the 21,310 instructors who taught in non-STEM fields, females outnumbered males (58% vs. 42%). Among instructors in STEM fields (*n* = 3,933), the proportion of males was nearly twice that of females (63% vs. 37%). Table 2 displays the distribution of instructors by gender and discipline group.

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 in the three summary measures, controlling for the influence of students' course motivation and work habits. This was followed by a 2 × 2 (Gender × Discipline Group) multivariate analyses of variance (MANOVA) on the reported frequencies of the seven TE teaching methods. Table 3 displays the descriptive statistics for the dependent variables and covariates.

Pearson *r* zero-order correlation coefficients among dependent variables and covariates are shown in Table 4. The strong correlations among the dependent variables justified the decision to conduct multivariate analyses.[3] Both covariates were significantly related to each of the three summary measures, which supports their inclusion in the analyses.

Given the large sample size in this study, we set the level of significance to .01. For all analyses, we computed partial eta squared (η^2_p) as a measure of effect size because it denotes proportion of variance explained in dependent variables. Univariate analyses were conducted following any significant multivariate effect, and the Bonferroni test was applied for post-hoc comparisons.

Table 2
*Number and Percentage of Female and Male Instructors in STEM and Non-STEM Fields (*N *= 25,243)*

	Female		Male	
Discipline category	*n*	%	*n*	%
STEM	1,450	37	2,483	63
Non-STEM	12,282	58	9,028	42

[2] STEM courses included science (agriculture, physical sciences, and biological sciences); technology (computer and information sciences); engineering (engineering, engineering technologies, and engineering-related fields); and mathematics (mathematics and statistics). Non-STEM courses included all other disciplines.

[3] An assumption of analysis of covariance is that the regression coefficients for the different groups are homogenous. To test this assumption, we compared the slopes for each of the three overall measures regressed on each of the two covariates between male and female instructors and between STEM and non-STEM instructors. There was no evidence that the slopes varied meaningfully by instructor gender or discipline group.

Table 3
Means and Standard Deviations of Student Ratings on Summary Measures, Teaching Methods, and Covariates by Instructor Gender in Non-STEM and STEM Groups

	Non-STEM		STEM		Total	
Source	*M*	*SD*	*M*	*SD*	*M*	*SD*
Summary measures						
PRO						
Female	4.15	0.43	3.99	0.47	4.13	0.44
Male	4.14	0.43	3.95	0.47	4.10	0.44
Excellence of teacher						
Female	4.21	0.61	4.06	0.68	4.19	0.62
Male	4.25	0.58	4.03	0.67	4.20	0.61
Excellence of course						
Female	4.07	0.56	3.84	0.58	4.05	0.57
Male	4.09	0.55	3.83	0.58	4.04	0.57
Teaching methods						
Displayed personal interest in students						
Female	4.40	0.48	4.25	0.55	4.39	0.49
Male	4.38	0.47	4.19	0.54	4.34	0.50
Helped students answer own questions						
Female	4.23	0.51	4.09	0.57	4.21	0.52
Male	4.22	0.49	4.04	0.55	4.18	0.51
Demonstrated importance of subject						
Female	4.40	0.45	4.20	0.52	4.38	0.46
Male	4.40	0.43	4.19	0.51	4.36	0.46
Made clear how topics fit						
Female	4.33	0.48	4.11	0.56	4.30	0.49
Male	4.33	0.47	4.09	0.53	4.27	0.49
Explained clearly and concisely						
Female	4.19	0.59	4.03	0.66	4.17	0.60
Male	4.21	0.55	3.97	0.65	4.16	0.59
Introduced stimulating ideas						
Female	4.24	0.52	3.94	0.61	4.21	0.53
Male	4.27	0.50	3.96	0.58	4.20	0.53
Inspired students to set high goals						
Female	4.12	0.54	3.80	0.60	4.08	0.55
Male	4.05	0.54	3.76	0.58	3.99	0.56
Covariates						
Course motivation						
Female	3.60	0.48	3.38	0.46	3.58	0.49
Male	3.52	0.47	3.43	0.45	3.50	0.47
Work habits						
Female	3.96	0.26	3.90	0.26	3.96	0.26
Male	3.94	0.26	3.89	0.26	3.93	0.26

Table 4

Correlation Coefficients for Relations Among Dependent Variables and Covariates

Measure	1	2	3	4	5
1. Progress on relevant objectives	-				
2. Excellence of teacher	.87	-			
3. Excellence of course	.88	.89	-		
4. Course motivation (covariate)	.43	.34	.54	-	
5. Work habits (covariate)	.38	.23	.31	.31	-

Note. All coefficients are significant at $p < .001$. $N = 25{,}243$.

Results

Main and Interaction Effects of Instructor Gender and Discipline Group on Student Ratings of Summary Measures of Teaching

In Research Question 1, we explored whether overall measures of teaching effectiveness differed by instructor gender and discipline group. Table 5 presents the results of the MANCOVA and subsequent univariate analyses. The main effects of both covariates were statistically significant and exhibited considerable effect sizes—$\eta^2_p = .35$ for course motivation and $\eta^2_p = .11$ for work habits—confirming their roles as important covariates. By multiplying those values by 100, we can interpret that course motivation and work habits explained about 35% and 11% of the variance, respectively, in the dependent variables. Table 6 provides adjusted mean scores and standard errors of the overall summary measures.

The multivariate interaction effect between instructor gender and discipline group was statistically significant, $F(3, 25235) = 15.23$, p < .001, but negligible ($\eta^2_p = .002$). The univariate analyses demonstrated that the interaction effect resided weakly in PRO ($\eta^2_p = .001$), excellence of the teacher ($\eta^2_p = .002$), and excellence of the course ($\eta^2_p = .002$). Thus the interaction explained no more than 0.2% of the variance in each dependent variable, which is trivial. As shown in Table 6, the effects of gender changed slightly at either level of discipline group. On excellence of the teacher, for example, in non-STEM courses students of male instructors assigned slightly higher ratings than those of female instructors, whereas the reverse was true in the STEM group: Women were rated slightly more highly than men. Still, the mean differences were small. Although the multivariate main effect of instructor gender was statistically significant, $F(3, 25235) = 18.88$, $p < .001$, the effect size was too small to indicate practical significance ($\eta^2_p = .002$). Moreover, none of the univariate tests on the individual dependent measures reached the .01 level of significance ($p = .09$ for PRO, $p = .10$ for excellence of teacher, and $p = .02$ for excellence of course). An examination of the marginal means for gender (see Table 3) reveals negligible differences between male and female instructors, ranging from 0.01 to 0.03 on a 5-point scale, suggesting students perceived neither gender to be superior in teaching to the other.

The MANCOVA revealed a significant main effect of discipline group on overall summary measures, $F(3, 25235) = 125.80$, $p < .001$, accounting for almost 2% of the variance in the dependent variables ($\eta^2_p = .015$). Univariate analyses indicated students tended to assign higher ratings to non-STEM instructors on all three summary measures than to STEM instructors. The marginal mean differences between non-STEM and STEM instructors ranged from 0.17 to 0.25, with the strongest effect found on excellence of the course ($M_{non\text{-}STEM} = 4.08$ and $M_{STEM} = 3.83$).

Table 5

Multivariate and Univariate Analyses of Covariance for Progress on Relevant Objectives (PRO), Excellence of Teacher, Excellence of Course

				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)	4475.71	< .001	.347	3621.97	<.001	.126	2211.59	<.001	.081	7966.34	<.001	.240
Work habits (covariate)	1060.49	< .001	.112	2168.25	<.001	.079	526.32	<.001	.020	760.26	<.001	.029
Gender	18.88	< .001	.002	2.96	.085	<.001	2.66	.103	<.001	5.68	.017	<.001
Discipline group	125.80	< .001	.015	193.82	<.001	.008	105.20	<.001	.004	262.87	<.001	.010
Gender × Discipline group	15.23	< .001	.002	32.02	<.001	.001	38.29	<.001	.002	45.54	<.001	.002

Note. Multivariate *F* ratios were generated from Pillai's statistic. [a]Multivariate *df* = 3, 25235. [b]Univariate *df* = 1, 25237.

Table 6
Adjusted Mean Scores and Standard Errors of Student Ratings Based on Course Motivation and Work Habits Grouped by Instructor Gender and Discipline Group

	Non-STEM		STEM	
Group	*M*	*SE*	*M*	*SE*
Progress on Relevant Objectives (PRO)				
Female	4.12	0.00	4.06	0.01
Male	4.15	0.00	4.01	0.01
Excellence of teacher				
Female	4.18	0.01	4.14	0.02
Male	4.26	0.01	4.09	0.01
Excellence of course				
Female	4.03	0.00	3.95	0.01
Male	4.11	0.00	3.91	0.01

Note. *N* = 25,243.

Main and Interaction Effects of Instructor Gender and Discipline Group on Student Ratings of the Seven Teaching Methods

Research Question 2 examined whether the frequencies with which instructors use the seven TE teaching methods, according to student reports, differed by instructor gender and discipline group. Table 7 displays the results of the MANOVA and subsequent univariate analyses.

The multivariate interaction effect was statistically significant, $F(7, 25233) = 19.59$, $p < .001$, but negligible ($\eta^2_p = .005$); weak univariate interaction effects were found in two of the seven teaching methods: helping students answer their own questions, $F(1, 25239) = 7.74$, $p < .01$, $\eta^2_p < .001$; and explaining material clearly and concisely, $F(1, 25239) = 7.74$, $p < .01$, $\eta^2_p = .001$. Similar patterns of interaction effects on the two teaching methods were identified: Although both genders employed either method with nearly identical frequency in non-STEM fields (i.e., $M_{male} = 4.22$ and $M_{female} = 4.23$ for helping students answer questions, and $M_{male} = 4.21$ and $M_{female} = 4.19$ for clarity and conciseness), women in STEM used the two methods slightly more often than did their male colleagues (i.e., for helping students answer their own questions, $M_{male} = 4.04$ and $M_{female} = 4.09$; and for clarity and conciseness, $M_{male} = 3.97$ and $M_{female} = 4.03$). However, as indicated by the minimal effect sizes, these differences are practically marginal.

The multivariate main effect for gender was statistically significant, $F(7, 25233) = 56.77$, $p < .001$, $\eta^2_p = .016$. Students gave female instructors slightly higher ratings than they gave male instructors on three teaching methods: displaying personal interest in students, $F(1, 25239) = 20.27$, $p < .001$, $\eta^2_p < .001$; helping students answer their own questions, $F(1, 25239) = 9.33$, $p < .01$, $\eta^2_p < .001$; and inspiring students to set and achieve challenging goals, $F(1, 25239) = 30.14$, $p < .001$, $\eta^2_p = .001$. However, in all cases the differences were trivial—the largest being on "inspiring students," where women were rated only slightly more highly than men ($M_{female} = 4.08$ and $M_{male} = 3.99$). The relative comparability between male and female instructors on student ratings of teaching methods helps to explain why no meaningful gender differences were found in overall summary measures.

The multivariate test on the main effect of discipline group was significant, $F(7, 25233) = 327.52$, $p < .001$, explaining approximately 8% of the variance ($\eta^2_p = .083$); students rated non-STEM instructors more highly on all seven TE teaching methods, with mean differences ranging from 0.17 to 0.32. The strongest effects were found on the methods of "introducing stimulating ideas about the topic" ($\eta^2_p = .039$) and "inspiring students to set and achieve challenging goals" ($\eta^2_p = .037$).

Table 7
Multivariate and Univariate Analyses of Variance for Student Ratings on Teaching Methods

			Univariate													
	Multivariate		Displayed personal interest in students		Helped students answer own questions		Demonstrated importance of subject		Made clear how topics fit		Explained clearly		Introduced stimulating ideas		Inspired students to set high goals	
Source	F^a	η^2_p	F^b	η^2_p	F^b	η^2_p	F^b	η^2_p	F^b	η^2_p	F^b	η^2_p	F^b	η^2_p	F^b	η^2_p
Gender (G)	56.77***	.016	20.27***	<.001	9.33**	<.001	.03	<.001	1.20	<.001	4.47	<.001	6.11	<.001	30.14***	.001
Discipline group (D)	327.52***	.083	411.60***	.016	317.81***	.012	661.00***	.026	684.71***	.026	360.29***	.014	1022.80***	.039	971.59	.037
G × D	19.59***	.005	5.62	<.001	7.74**	<.001	.69	<.001	.64	<.001	17.82***	<.001	.03	<.001	.43	<.001

Note. Multivariate *F* ratios were generated from Pillai's statistic. [a]Multivariate *df* = 7, 25233. [b]Univariate *df* = 1, 25239.
** *p* < .01. *** *p* < .001

Discussion

The results of the current study can be summarized as follows. First, instructor gender has no practically meaningful effects on student ratings of either overall summary measures or instructor use of teaching methods, in both STEM or non-STEM fields. Students rated their overall progress, the quality of the teacher and course, and the frequency of teaching methods very similarly regardless of whether they were taught by a man or a woman. Second, non-STEM instructors tend to receive higher student ratings than their STEM peers on overall summary measures and use of effective teaching methods. Third, course motivation and work habits are 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 supports the results of previous large-scale research (Centra, 2009; Centra & Gaubatz, 2000; Feldman, 1993; Smith et al., 2007). Also, women and men do not differ practically when using the essential teaching methods that facilitate student learning. The exceptionally weak interaction effects found between instructor gender and discipline group suggest that in the natural settings, (a) neither gender is superior in teaching, in either STEM or non-STEM fields; and (b) neither gender exhibits a stronger preference for particular teaching methods, whether teaching STEM or non-STEM courses.

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 stereotypes may effectively reduce potential bias associated with student expectations.

Results of the study also suggest that gender bias in student ratings as found in previous research (e.g., Boring, 2017; Boring, Ottoboni, & Stark, 2016; MacNell, Driscoll, & Hunt, 2014; for a critique, see Benton & Li, 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 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, faculty evaluation and consequent personnel decisions are prone to biases inherent in individual and collective perceptions and expectations of certain demographic or cultural groups at various stages of the process. 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*.

The strongest effects observed in the current study were found between STEM and non-STEM fields. Students gave non-STEM instructors higher ratings on all summary measures, especially on excellence of the course, and on all teaching methods. This indirectly supports the notion that what teachers do in the classroom is connected to overall evaluations. Previous research found that all seven TE methods are highly correlated with overall ratings of the teacher and the course (Benton et al., 2015). The two teaching methods where the greatest differences were found between the discipline groups, "introducing stimulating ideas about the topic" and "inspiring students to set and achieve challenging goals," are positively correlated with student progress ratings on eight and ten of 12 learning objectives, respectively (Benton et al., 2015). Ratings on the behaviors teachers exhibited in the classroom corresponded with students' overall impressions of their own progress on the learning objectives, how well they believed they were taught, and their overall impressions of the course. If higher ratings were not accompanied by greater use of effective teaching methods, there would be cause for concern.

Because we controlled for student motivation and work habits, it is less likely that student characteristics alone can explain the differences between STEM and non-STEM results. However, there remain several other explanations—beyond teaching methods—of why students assign higher ratings in non-STEM courses. Students may perceive STEM courses as more difficult, and they may lack the necessary foundation in quantitative skills. Moreover, the hierarchical structure of STEM content makes prerequisite knowledge essential for success. Students who are deficient in such knowledge therefore tend to assign lower ratings (Benton et al., 2015). Finally, relative to non-STEM faculty, those in STEM may place a higher priority on research than teaching, thereby devoting more time and effort to the former than to the latter. Although the corresponding differences are trivial, they may suggest a connection between the behaviors teachers exhibit in the classroom and students' overall ratings.

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. Fourth, correlational methods employed in this study did not establish a cause-effect relationship between student ratings of teaching methods and overall summary measures.

Implications

The effects of instructor gender and its interaction with academic-discipline group do not exert much influence on overall IDEA SRI measures. The most telling difference in ratings is observed not between men and women but between STEM and non-STEM instructors. 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.

References

Allport, G. W. (1954). The nature of prejudice. Cambridage, 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. 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), 170–10. 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 from http://www.ideaedu.org/whats-in-the-study-exposing-validity-threats-in-the-macnell-driscoll-and-hunt-study-of-gender-bias/

Benton, S. L., Gross, A., & Brown, R. (2012). Which learning outcomes and teaching methods are instructors really emphasizing in STEM courses? Presented at the American Association of Colleges and Universities Network for Academic Renewal, Kansas City, MO.

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.

Biglan, A. (1973). Relationships between subject matter characteristics and the structure and output of university departments. *Journal of Applied Psychology, 57*(3), 204–213.

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

Braskamp, L. A., & Ory, J. C. (1994). Assessing faculty work: Enhancing individual and institutional performance. San Francisco: Jossey-Bass.

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.

Cashin, W. E. (1990). Students do rate different academic fields differently. *New Directions for Teaching and Learning, 1990*(43), 113–121. http://doi.org/10.1002/tl.37219904310

Centra, J. A. (1993). Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness. San Francisco: Jossey-Bass.

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. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education, 9*(3), 199–242–45. http://doi.org/10.1007/BF00976997

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.

Franklin, J., & Theall, M. (1992). Disciplinary differences: Instructional goals and activities, measures of student performance, and student ratings of instruction. Presented at the Annual Conference of the American Educational Research Association, San Francisco, California.

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

Kember, D., & Leung, D. Y. P. (2011). Disciplinary differences in student ratings of teaching quality. *Research in Higher Education, 52*(3), 278–299. http://doi.org/10.1007/s11162-010-9194-z

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

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

Marsh, H. W., & Dunkin, M. J. (1997). Students’ evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241–320). New York: Agathon Press. http://doi.org/10.1007/1-4020-5742-3_9

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 (0.5.1) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/gender/index.html

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

Sixbury, G. R., & Cashin, W. E. (1995). *IDEA Technical Report No. 9: Description of database for the IDEA Diagnostic Form*. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

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.*

T: 800.255.2757
T: 785.320.2400

301 South Fourth St., Suite 200
Manhattan, KS 66502-6209
E: info@IDEAedu.org
IDEAedu.org