frontiers in Education

ORIGINAL RESEARCH
published: 11 January 2021
doi: 10.3389/feduc.2020.571287

Gender Stereotypes in Student Evaluations of Teaching

Emma A. Renström[1]*, *Marie Gustafsson Sendén*[2] and *Anna Lindqvist*[3]

[1] Department of Psychology, University of Gothenburg, Gothenburg, Sweden, [2] Department of Psychology, Stockholm University, Stockholm, Sweden, [3] Department of Psychology, Lund University, Lund, Sweden

This paper tests how gender stereotypes may result in biased student evaluations of teaching (SET). We thereby contribute to an ongoing discussion about the validity and use of SET in academia. According to social psychological theory, gender biases in SET may occur because of a lack of fit between gender stereotypes, and the professional roles individuals engage in. A lack of fit often leads to more negative evaluations. Given that the role as a lecturer is associated with masculinity, women might suffer from biased SET because gender stereotypes indicate that they do not fit with this role. In two 2 × 2 between groups online experiments (*N*'s = 400 and 452), participants read about a fictitious woman or man lecturer, described in terms of stereotypically feminine or masculine behavior, and evaluated the lecturer on different SET outcomes. Results showed that women lecturers were not disfavored in general, but that described feminine or masculine behaviors led to gendered evaluations of the lecturer. The results were especially pronounced in Experiment 2 where a lecturer described as displaying feminine behaviors was expected to also be more approachable, was better liked and the students rather attended their course. However, a lecturer displaying masculine behaviors were instead perceived as being more competent, a better pedagogue and leader. Gender incongruent behavior was therefore not sanctioned by lower SET. The results still support that SET should not be used as sole indicators of pedagogic ability of a lecturer for promotion and hiring decisions because they may be gender-biased.

Keywords: student evaluations of teaching (SET), gender stereotypes, gender bias, social psychology, experiment

OPEN ACCESS

Edited by:
Jaime Ibáñez Quintana, University of Burgos, Spain

Reviewed by:
Manpreet Kaur Bagga, Partap College of Education, India
Alan Garnham, University of Sussex, United Kingdom

***Correspondence:**
Emma A. Renström
emma.renstrom@psy.gu.se

Specialty section:
This article was submitted to Teacher Education, a section of the journal Frontiers in Education

Received: *10 June 2020*
Accepted: *07 December 2020*
Published: *11 January 2021*

Citation:
Renström EA, Gustafsson Sendén M and Lindqvist A (2021) Gender Stereotypes in Student Evaluations of Teaching. Front. Educ. 5:571287. doi: 10.3389/feduc.2020.571287

INTRODUCTION

The purpose of this article was to test the impact of gender stereotypes in student evaluations of teaching (SET), in two online social psychological experiments. Previous research in this field indicates a gender bias in SET where women generally receive lower SET compared to men (e.g., MacNell et al., 2015; Boring, 2016; Mengel et al., 2018; Mitchell and Martin, 2018; Fan et al., 2019). With this article, we contribute to an ongoing discussion about the use of SET, both as formative and summative evaluations of teaching and teachers. We provide new insights into the mechanisms behind SET and how they relate to a lecturer's gender identity and gendered behavior.

Taking a social psychological perspective, gender biases may occur because gender stereotypes prescribe and proscribe certain behaviors for individuals of different genders. Specifically, when gender stereotypes and professional roles do not fit, the individual can be sanctioned with negative evaluations (Heilman, 2001; Heilman and Chen, 2005; Heilman and Haynes, 2005). In this article, we test to what extent women lecturers in higher education are sanctioned by low SET due to a tradeoff between behaviors expected from the supposedly masculine-coded role as a university lecturer, and the stereotypes about how women should and should not be.

Student Evaluations of Teaching

Originally, SET were introduced for formative purposes. That is, the evaluations were to be used in order to improve and shape the quality of teaching (Hornstein, 2016). Since then, SET has become a primary indicator of summative evaluations of a lecturer's performance. That is, SET are used as an overall sum of pedagogical competence, often as the sole indicator of this competence (Berk, 2005; Galbraith et al., 2012; Spooren et al., 2013). SET are now often used for promotion and hiring decisions (Cashin, 1999; Seldin, 1999; Clayson, 2009; Davis, 2009; Seldin et al., 2010), indicating that it is important to understand systematic variations in SET.

SET were first criticized by Adams (1997), where he pointed out several flaws such as validity, reliability, gender bias, and a number of other related issues (Yunker and Yunker, 2003; Wright, 2006; Beecham, 2009; Hoefer et al., 2012; Spooren et al., 2013; Braga et al., 2014; Stark and Freishtat, 2014; Boring et al., 2016). It is suggested that SET mainly reflects satisfaction with teaching among students after they have finished a course. As such, it is argued that SET rather should be seen as a popularity measurement, rather than a measurement of teaching capability (Beecham, 2009; Spooren et al., 2013; Braga et al., 2014; Stark and Freishtat, 2014). This paves the way for both individual and contextual factors to exert influence regarding high or low evaluations and leads to the aim of the present article—to test if gender stereotypes influence SET.

Several studies have shown a gender bias in SET, although the results are inconclusive. Many studies have shown that women receive lower evaluations than men (MacNell et al., 2015; Boring et al., 2016; Mengel et al., 2018; Mitchell and Martin, 2018). For instance, Boring et al. (2016) showed a systematic gender bias in SET where women lecturers received lower evaluations on seemingly objective aspects, such as how promptly assignments were graded. Likewise, Mitchell and Martin (2018) showed that a woman lecturer was rated lower on other similar aspects, such as the course itself, work load, the technology, etc. However, some studies show that women receive higher ratings than men (Rowden and Carlson, 1996; Bachen et al., 1999), and finally, some have not found a difference between evaluations of women and men (Feldman, 1993; Centra and Gaubatz, 2000). These results imply that gender of a lecturer alone is not sufficient to explain variations in SET between women and men lecturers. One possible cause to the inconsistencies in earlier results may be that both individual and contextual factors interact with a lecturer's gender (Boring et al., 2016). For instance, Boring et al. (2016) found that the gender bias in SET varied with, for example, discipline. These results are supported by Mengel et al. (2018), who showed that the gender bias is magnified in mathematical courses, and particularly pronounced for younger women lecturers. One explanation might be that the STEM-field (Science, Technology, Engineering, and Math) is heavily dominated by men (Makarova et al., 2019), where (younger) women accordingly violate the gender norms, resulting in a lack of fit between the expectations of their gender role and the expectations of the role as a university lecturer, which could explain the bias (Heilman, 1983, 2012). Such lack of fit, described more below, indicate that a woman lecturer behaving in a "masculine" way may receive different SET as compared to a woman lecturer acting in a "feminine" way, which essentially decreases the lack of fit. To better understand the complexity of how gender, stereotypes and fit between a lecturer's gender and their behavior operate to influence biases in SET, we now turn to social psychological theory.

Gender Stereotypes

Gender stereotypes are collective mental representations about what is typical regarding women and men when it comes to personality, behavior, and/or expression (Ellemers, 2018). This means that gender stereotypes are shared generalizations about women and men, and the consensus of these generalizations among the population is high (Hentschel et al., 2019). The content of the gender stereotypes pertain to two core dimensions in social judgment, referred to as agency and communion (Abele and Wojciszke, 2014). Agency refers to goal-achievement, whereas communion refers to the maintenance of social relationships (Bakan, 1966). Women are more often perceived as communal (e.g., caring, sensitive, loyal, and understanding; Eagly and Wood, 2012), while men are more often perceived as agentic (e.g., independent, assertive, dominant, self-reliant, and determined). Hence, agentic traits are traditionally associated with masculinity, while communal traits are traditionally associated with femininity. Importantly, gender stereotypes function both prescriptively (what women and men should engage in, and how they should be), and proscriptively (what they should not engage in and be) (Gustafsson Sendén et al., 2019; Hentschel et al., 2019).

When gender stereotypes are fulfilled, that is, when women perform communal tasks and men perform agentic tasks, individuals are positively evaluated. Thus, lecturers who adhere to gendered expectations can be evaluated more favorably (Andersen and Miller, 1997). For example, Boring (2016) found that women lecturers received the highest ratings on availability and quality of contact—two characteristics typical of the stereotypes for women (Abele and Wojciszke, 2014). In relation to social perception and evaluation of others, the problem with stereotypes becomes evident when they are challenged—when gender and role, or behavior, mismatch. When stereotypes regarding roles or behavior and gender are incongruent (i.e., lack of fit), individuals are likely to be sanctioned and negatively evaluated (Heilman, 1983, 2012; Eagly and Karau, 2002; Heilman and Okimoto, 2007; Brescoll et al., 2010). Rudman et al. (2012) discuss a gender backlash effect where women can reach higher positions through agentic behaviors, but they are at the same time disliked and hence not viewed as hirable. This leads women to a situation where they are forced between being liked or being respected, which undermines their ability to achieve positions of power (Rudman et al., 2012). For instance, when women engage in behaviors typically considered as masculine, they are less liked and their behavior is found to be less socially accepted, as compared to when men engage in the same behavior (Bartol and Butterfield, 1976; Jago and Vroom, 1982; Carli, 1990; Carli et al., 1995; Heilman and Okimoto, 2007). This seems to be true in students' perceptions of lecturers as well. When gender roles are violated by lecturers, students become critical

(Chamberlin and Hickey, 2001; Sprague and Massoni, 2005). This suggests that if gender stereotypes are responsible for the variation in SET between women and men lecturers that has been observed in previous research, the role as a lecturer is coded as masculine. Traditionally, higher education has been exclusively for men, which could still affect how the role as a university lecturer is perceived in terms of gender. Moreover, being a lecturer at a higher education institution is a leadership role, and because leadership and authority traditionally are associated with masculinity (see Heilman and Okimoto, 2007), women lecturers violate gender stereotypes and may face biases and criticism (Eagly and Karau, 2002). Hence, women lecturers must balance the demands of their gender role, as well as the demands of being an authority figure, which inevitably will lead to some sort of discrepancy. Taken together, theory and empirical studies highlight the difficulty that women lecturers have in balancing the tension between agentic demands from the leadership role and communal demands from the gender role (Zhen et al., 2018).

Overview of the Present Research

The present research zooms in on the discrepancy between gender stereotypes and the role as a university lecturer as a source of gender bias in SET. Specifically, we test if women lecturers are sanctioned if they do not engage in traditionally feminine behaviors, or lack traditionally feminine characteristics (Rudman, 1998; Rudman and Glick, 2001). The following hypotheses are formulated:

H1: Women lecturers receive lower SET on average, compared to men lecturers.

H2: A woman lecturer described as having traditionally masculine behavior and characteristics, receive the lowest SET.

In two experiments, students were presented with a description of a fictive lecturer. The descriptions varied with respect to the lecturer's gender (the lecturer was referred to as either "she" or "he" in the text). Moreover, the behavior and characteristics of the lecturer were described as either stereotypically feminine or stereotypically masculine. In Experiment 1, the description of the lecturer contained both positive and negative feminine/masculine behaviors and traits. In Experiment 2, the valence of feminine/masculine behaviors and traits (i.e., positive and negative) was even more balanced. Participants' task was to rate the lecturer on common SET items. Experiment 1 used a wide range of SET items, mainly from previous literature. In Experiment 2, the number of items were reduced due to semantical overlap.

The studies were carried out in accordance with the national guidelines on ethical research established by the Swedish Research Council retrievable at: https://publikationer.vr.se/en/product/good-research-practice/.

EXPERIMENT 1

Because our hypotheses are formulated to test the potential mismatch between the role as a university lecturer, and the female gender role, we first established that the role as a university lecturer was indeed coded as masculine. In a pilot study, 82 students read a description of a lecturer. The description varied with respect to gender stereotypical (feminine and masculine) characteristics and behaviors of the lecturer, but no actual gender information was provided (i.e., we replaced the pronoun with X). After reading the description of the lecturer, participants indicated what gender they thought the lecturer had, as a free-text response. Across the feminine ($n = 33$) and masculine ($n = 49$) conditions, 74 (90%) participants indicated that the lecturer was a man, only 8 (10%) indicated a woman (masculine condition: man = 44, woman = 5; feminine condition: man = 30, woman = 3). No other genders were suggested. Hence, the role as university lecturer is clearly associated with masculinity.

Method

Participants, Design, and Procedure

Four hundred US students, who were currently enrolled in higher education, were recruited from the platform *Prolific Academic*. Participant gender was assessed by free-text (Lindqvist et al., 2020); the sample consisted of 196 men (49%), 185 women (46%), 21 participants (5%) gave another response than woman/man.[1] Mean age was 27 years old (range: 18–63, $SD = 8.26$).

To assess the impact of lack of fit between the lecturer role and gender role, we designed an experiment where the lecturer's gender and behavior varied between conditions. The design was a 2 (gender: she/he) × 2 (behavior: feminine/masculine), between groups factorial design. For example, in the feminine version, the lecturer was described as supportive and caring, being available for students, being responsive and empathic, while the masculine version was described as more focused on the research, being assertive and demanding, expecting hard work, and being unavailable. The descriptions were balanced in that the feminine version also contained some negative feminine traits, such as being uncertain, whereas the masculine version contained some positive masculine traits, such as being certain. The descriptions are provided in the **Supplementary Material**. Participants were randomly assigned to one of the four conditions (n's= she/masculine = 119, she/feminine = 89, he/masculine = 99, he/feminine = 94).

Measures

To measure SET, a range of measures from previous research were included. The Professor Effectiveness scale (Goebel and Cashen, 1979; Wilson et al., 2014), The Brief Professor-Student Rapport Scale (Ryan and Wilson, 2014) with two sub-scales (Perceptions of the teacher and Student Engagement). Personal characteristics of the lecturer were assessed by items suggested by MacNell et al. (2015) and Boring (2016). To assess perceptions of the lecturer's competence, we included items referring to more general perceptions of the course and the pedagogy, since these may better reflect competence compared to the evaluation of individual characteristics. These items were averaged into a mean index. Two items measured the difficulty level of the course, and two items measured the general impression of the course. Finally, participants rated warmth and competence (Fiske et al., 2007). Where indices were made of the scales, we averaged the items into

[1] 3 did not respond at all, 2 agender, 13 non-binary, 1 trans male and 2 put two-spirit.

TABLE 1 | Scales and items used in the experiments.

Scale	Items	Responses	Cronbach's α
Professor effectiveness (Goebel and Cashen, 1979; Wilson et al., 2014)		1 = Strongly disagree 7 = Strongly agree	Analyzed separately
	The lecturer encourages questions		
	The lecturer expects good work		
	The lecturer assigns too much work		
	The lecturer is organized		
	The lecturer can explain concepts		
	The lecturer behaves in a friendly manner		
	The lecturer is generally a good teacher		
The Brief Professor-Student Rapport Scale (Ryan and Wilson, 2014)		1 = Strongly disagree 7 = Strongly agree	
	The lecturer is compassionate		
	The lecturer is enthusiastic		
	The lecturer is reliable		
	The lecturer is receptive		
	The lecturer cares about the class		
	The lecturer encourages questions and comments from students		
	The lecturer makes class enjoyable		
Perceptions of the teacher			0.88
Student engagement			0.91
MacNell et al. (2015)		1 = Strongly disagree 7 = Strongly agree	0.92
	The lecturer is caring		
	The lecturer is consistent		
	The lecturer is enthusiastic		
	The lecturer is fair		
	The lecturer is helpful		
	The lecturer is knowledgeable		
	The lecturer is professional		
	The lecturer is prompt		
	The lecturer is respectful		
	The lecturer provides praise		
	The lecturer provides feedback		
Boring (2016)		1 = insufficient, 2 = average, 3 = good and 4 = excellent	Analyzed separately
	The lecturer's preparation and organization of classes		
	The quality of the instructional material		
	The lecturer's ability to encourage work		
	The lecturer's availability		
	The quality of contact		
	The lecturer's ability to lead the class		
	The lecturer's ability to relate to current issues		
	The lecturer's contribution to the students' intellectual development		
Pedagogy items		1 = Not at all, 7 = Very much	0.87
	The content of the course aligns with the learning outcomes of the course		
	The course offers opportunities to lean and understand the content of the course		
	Different modules of the course are integrated with each other		
	The examinations on the course measures the learning outcomes		
	Do you think that the students on the course have learnt much compared to what they knew before the course		
	Do you think the requirements for the grading have been clearly communicated		
Difficulty level			Analyzed separately

(Continued)

TABLE 1 | Continued

Scale	Items	Responses	Cronbach's α
	How many hours do you think that the students at the course study	Responses were made as free text[a]	
	What is the level of requirement	1 = Extremely easy 7 = Extremely difficult	
Single-items			Analyzed separately
	What is your overall impression of the course	1 = Extremely bad 7 = Extremely good.	
	How interested would you be in attending a course with the lecturer	1 = not at all interested, 7 = Very interested	
			Analyzed separately
Fiske et al. (2007)		1 = Strongly disagree 7 = Strongly agree	
	Warmth		
	Competence		

[a] *This item was re-formulated in Study 2 since some participants expressed that it was difficult to understand.*
Perceptions of the teacher and Student engagement are subscales of The Brief Student-Rapport Scale.

a mean index. Cronbach's α's for these scales are shown in **Table 1**, where it is also detailed if the items were analyzed separately (i.e., not included in a scale). The questions are summarized in **Table 1**.

Results

For all of the outcome measures detailed in **Table 1**, we computed 2 × 2 ANOVAs with gender of the lecturer (she/he) and gendered behavior (feminine/masculine) as between-participant factors. We also included participant gender as covariate. Means, standard deviations and *F*-values for the main effects are shown in **Table 2**. Only the main effects are presented, because none of the interaction effects were significant.

The first hypothesis stated that women lecturers overall should receive lower SET than men. The results showed no main effects of the lecturer's gender on any of the outcome variables, see **Table 2**. The second hypothesis stated that women lecturers described as having masculine characteristics and behavior should receive the lowest SET. This hypothesis implies that we would see interaction effects between gender of the lecturer and described behavior. However, none of the interactions were significant. Thus, the results indicate that there were no differences between how a woman lecturer was rated depending on feminine/masculine behavior, as compared to a man lecturer described with feminine/masculine behavior. This means that neither of the hypotheses were supported. Interestingly, there were significant main effects of whether the lecturer was described as having feminine or masculine characteristics on all outcome variables. The means are shown in **Table 2**. For easier overview, significant differences in favor of the feminine description are marked in bold, while differences in favor of the masculine description are marked in gray.

In sum, participants rated a feminine behavior more positively than the masculine behavior on almost all the outcome measures. The difference on many items are unsurprising since the text in the feminine condition described a lecturer that was more involved with the students and teaching, therefore it can be expected that students would prefer a lecturer with these characteristics. For instance, in the Professor Effectiveness scale, the items *encourages questions, is organized, can explain concepts, behaves in a friendly manner*, and *is generally a good teacher* should receive higher values based on the text in the feminine condition. An interesting finding was that the participants expected that the masculine lecturer would *expect good work* and *assign too much work* to a higher degree compared to the feminine lecturer. Other results that are not easily explained by the descriptions of the lecturer are the items related to difficulty. The participants thought that the course had higher requirements and that students at the course studied more when the behavior of the lecturer was masculine.

Combined, the results indicate that the participants rate a lecturer described in feminine terms more positively, and they rather attend their course, compared to a lecturer described in masculine terms. However, the participants thought that the masculine behavior implied higher demands and a more difficult course, where students actually did put in more hours. These are not unambiguously negative features from a learning perspective.

Finally, the lecturer with masculine behavior was rated as less competent than the lecturer with feminine behavior. Even though the effect was smaller compared to the other effects in this study, it was significant. This was surprising since competence has been strongly associated with masculinity (Fiske et al., 2007). However, recent research show that competence is one aspect of gender stereotypes that has changed the most over the years, and that women now sometimes are perceived as more competent than men (Gustafsson Sendén et al., 2019; Eagly et al., 2020). Hence, the results are not contradicting of recent research. Also, in the masculine condition, the lecturer was described as more competent as a researcher than teacher, while the feminine behavior was described as more competent in

TABLE 2 | Means, standard deviations (in parentheses) and *F*-values from univariate ANOVAs for main effects of conditions (she/he; feminine/masculine), in Experiment 1, $N = 400$.

	Condition					
	Lecturer gender			Description		
Outcome	**She**	**He**	***F***	**Feminine**	**Masculine**	***F***
Professor effectiveness						
Encourages questions	3.88 (2.39)	4.34 (2.33)	1.65	**6.25 (0.95)**	**2.35 (1.63)**	751.39***
Expects good work	6.10 (1.00)	5.96 (0.85)	0.52	5.61 (0.94)	6.38 (0.83)	67.13***
Assigns too much work	3.87 (1.28)	3.85 (1.36)	0.28	3.17 (1.18)	4.42 (1.14)	109.38***
Is organized	5.55 (1.23)	5.47 (1.15)	0.69	**5.66 (1.18)**	**5.39 (1.19)**	4.89*
Can explain concepts	5.01 (1.70)	5.24 (1.60)	0.26	**6.25 (0.93)**	**4.20 (1.54)**	224.55***
Behaves in a friendly manner	4.60 (1.94)	4.82 (1.93)	0.01	**6.33 (0.95)**	**3.39 (1.47)**	505.21***
Is generally a good teacher	5.00 (1.73)	5.22 (1.65)	0.05	**6.43 (0.79)**	**4.03 (1.15)**	369.13***
Professor-student Rapport scale						
Perceptions of the teacher	4.95 (1.30)	5.03 (1.40)	0.40	**6.16 (0.71)**	**4.05 (0.94)**	582.12***
Student engagement	4.53 (1.68)	4.84 (1.71)	1.13	**6.22 (0.72)**	**3.42 (1.14)**	763.92***
MacNell	5.08 (1.06)	5.13 (1.12)	0.43	**6.00 (0.62)**	**4.39 (0.83)**	432.98***
Boring (scale 1–4)						
Preparation and organization	3.15 (0.82)	3.08 (0.80)	1.66	**3.39 (0.63)**	**2.90 (0.87)**	38.25***
Quality of instructional material	3.03 (0.87)	3.01 (0.83)	0.44	**3.28 (0.69)**	**2.80 (0.91)**	31.92***
Ability to encourage work	2.67 (0.97)	2.77 (1.01)	0.05	**3.36 (0.72)**	**2.20 (0.87)**	194.20***
Availability	2.57 (1.15)	2.66 (1.14)	0.04	**3.52 (0.67)**	**1.88 (0.89)**	397.82***
Quality of contact	2.48 (1.19)	2.65 (1.20)	0.27	**3.53 (0.65)**	**1.77 (0.92)**	437.75***
Ability to lead the class	2.78 (1.00)	2.90 (0.96)	0.05	**3.49 (0.61)**	**2.31 (0.91)**	202.50***
Ability to relate to current issues	2.72 (0.92)	2.69 (0.95)	1.03	**3.17 (0.76)**	**2.33 (0.92)**	91.65***
Contribution to the students' intellectual development	2.78 (1.08)	2.84 (1.02)	0.24	**3.50 (0.67)**	**2.24 (0.96)**	207.31***
Pedagogy index	5.01 (1.25)	5.05 (1.16)	0.56	**5.76 (0.85)**	**4.44 (1.13)**	160.61***
Difficulty level						
How many hours do you think the student study?	9.94 (7.61)	8.81 (6.57)	2.03	7.94 (6.43)	10.60 (7.50)	12.96***
What is the level of requirement?	5.05 (1.21)	4.85 (1.25)	1.12	4.36 (1.12)	5.44 (1.10)	87.50***
Single-items						
Overall impression of the course	4.58 (1.66)	4.63 (1.56)	1.02	**5.75 (1.00)**	**3.67 (1.41)**	265.69***
Would you like to attend a course with the lecturer	4.17 (2.22)	4.44 (2.08)	0.00	**5.96 (1.16)**	**2.98 (1.83)**	325.85***
Fiske et al., 2007						
Warmth	4.09 (1.88)	4.38 (1.96)	0.45	**5.82 (1.09)**	**2.93 (1.39)**	484.83***
Competence	5.98 (1.08)	5.91 (1.13)	0.92	**6.23 (0.90)**	**5.70 (1.21)**	23.07***

****p* < 0.001, ***p* < 0.01, **p* < 0.05.

Bold figures indicate significant differences in favor of a woman/feminine lecturer, gray highlighting indicate significant differences in favor of a masculine lecturer.

pedagogy. It is possible that this asymmetry between competence in different areas influenced the participants when they made the overall competence rating. From a student perspective, pedagogical competence should be more important in SET than research competence.

One reason for the lack of main effects of the lecturer's gender, or interactions with description of behavior and characteristics, may be that the feminine version overall was seen as more positive from a student's perspective. Hence, in a second experiment, the descriptions of the lecturer were more ambiguous, so that the feminine condition also entailed more negative feminine traits and the masculine condition entailed more positive masculine traits. We also reduced the number of outcome variables, and focused on assessments of the course that were not directly related to the individual described.

EXPERIMENT 2

Methods

Participants, Design, and Procedure

We recruited 452 US students (149 from *Prolific Academic* and 303 from *M-turk*). The participants were self-defined as 143 men (32%), 241 women (53%), 58 (15%) gave another response than woman/man.[2] Mean age was 25 years (range: 18–65, $SD = 6.43$).

The design was the same as in Experiment 1, that is a 2 (gender of lecturer: she/he) × 2 (description: feminine/masculine), between groups factorial design. Participants were randomly assigned to one of the four conditions (*n's* = she/masculine = 112, she/feminine = 100, he/masculine = 122, he/feminine =

[2]63 did not respond at all, 1 agender, 3 non-binary, 1 trans femme.

118). As mentioned, the feminine and masculine descriptions were now more balanced with respect to valence of described traits and behaviors. For instance, the feminine description detailed that the lecturer appeared afraid of students if being criticized, and problems in the teaching team where the lecturer lacked leadership skills and confidence (Abele and Wojciszke, 2014). Because we still kept the positive aspects in the description, such as being considerate, sympathetic and caring, the description was ambivalent on purpose. The masculine description underwent the same procedure, where that the lecturer was described as confident and convincing, ambitious, competent and professional, and that these traits were applied not only to research but also to teaching. By keeping some of the negative aspects from the previous description, such as being seen as unapproachable, research focused and rigid, this description also became ambivalent on purpose.

Measures

The outcome measures assessed pedagogy and evaluations of the course, rather than traits of the lecturer. The pedagogy items formed a scale with a mean index and were the same as in Experiment 1 ($\alpha = 0.85$). The items measuring difficulty level of the course were also the same, except for the item measuring perceived amount of study hours. This time, perceived amount of study hours was assessed with a scale from *1 = Very little time* to *7 = Very much time* instead of a free-text response, to make it possible to include the item in the mean index of difficulty level, instead of analyzing it separately. We kept the item "The lecturer assigns too much work" from the Professor effectiveness scale (Goebel and Cashen, 1979; Wilson et al., 2014) as it fitted nicely with the other difficulty level items. These three items were averaged into a mean index, $\alpha = 0.70$. Also, the single items regarding overall impression of and interest in attending the course were the same as in Experiment 1. We added 2 items of general impression: What is your overall impression of the lecturer? and How does the lecturer seem to be as a leader of the teaching team? Answers ranged from *1 = Extremely bad* to *7 = Extremely good*. Three items asked about specific traits and engagement: Do you think of the lecturer as a serious person? Do you think that the lecturer is knowledgeable? and Do you think that the lecturer is engaged in the teaching? Answers ranged from *1 = No, not at all* to *7 = Yes, definitely*. We also kept the item measuring competence and "What is your impression that the students think of the lecturer?" Finally, we kept the questions by Boring et al. (2016) because they focused more on the lecturer's ability than individual traits (see **Table 1**).

Results

For all outcome measures, we computed 2 × 2 ANOVAs with gender of the lecturer (she/he) and description (feminine/masculine) as between-participants factors. Participant gender was again included as covariate. Means, standard deviations and *F*-values for the main effects are shown in **Table 3**. Only the main effects are included, because none of the interaction effects were significant. For easier overview, we again marked significant differences in favor of the feminine lecturer (or a woman lecturer) in bold, while differences in favor of the masculine lecturer is marked in gray.

Table 3, shows a general pattern were type of behavior is significant on most outcome variables. For some outcomes, gender of the lecturer (she/he) was significant.

The first hypothesis stated that women should receive lower SET on average, compared to men. In contrast to Hypothesis 1, the effects were rather in favor of the woman. For instance, the overall impression of the course was higher for the woman, and she was also rated as better at pedagogy, compared to the man. Three items in the Boring (2016) scale were also significant in favor of a woman lecturer: preparation and organization, ability to relate to current issues and contribution to the students' intellectual development, which at least partly aligns with Boring's results. However, it should be noted that the effects were rather weak.

The second hypothesis focused on the interaction between gender of the lecturer (she/he) and description of behavior and characteristics (feminine/masculine), where we expected that a masculine woman would be rated lowest on SET. Because no interactions were significant, H2 was not supported. Hence, the results so far are largely in line with the results found in Experiment 1. This means that gender incongruent behavior, neither for women nor men lecturers, seem to lead to lower SET.

Similar to Experiment 1, there were several main effects of description (i.e., feminine/masculine). However, in contrast to Experiment 1, the effects were not consistently in favor of the feminine behavior, which indicate that we managed to make the descriptions more ambiguous. First, the masculine behavior seemed to reflect perceptions of being a better pedagogue. The feminine behavior was seen as better when it comes to encouraging work, being available, better quality of contact and better at relating to current issues—again largely in line with Experiment 1 and Boring (2016), and also in line with a feminine gender stereotype (Abele and Wojciszke, 2014). As in Experiment 1, the masculine behavior was perceived as "tougher," such that ratings of the lecturer described as masculine were higher on difficulty as compared to the feminine condition.

The masculine behavior was perceived as conforming to traditional male stereotypes of leadership and competence, such that the lecturer was seen as more serious, knowledgeable and competent, as well as being a better leader of the teaching team and the class. A possible reason for the shift in competence from the feminine behavior in Experiment 1 to the masculine behavior in Experiment 2 is most likely due to that the masculine description this time contained having the competence to, for instance, respond to students' questions and being more involved in the course in general.

While the participants rated masculine behavior higher on pedagogy, leadership, and learning, they still preferred the lecturer with the feminine behavior. The feminine behavior was rated higher on overall impression, and engagement in teaching. The students rated feminine behavior as more liked, and they expressed more interest in attending a course with a lecturer acting more feminine rather than masculine. Other stereotypically feminine characteristics that was rated higher in the feminine condition was ability to encourage work and

TABLE 3 | Means, standard deviations (in parentheses) and *F*-values in Study 2 ($N = 452$).

	Condition					
	Lecturer gender			Description		
Outcome	**She**	**He**	***F***	**Feminine**	**Masculine**	***F***
Pedagogy items (scale 1–7)						
	5.34 (0.92)	**5.12 (1.01)**	4.76*	5.12 (0.91)	5.33 (0.97)	3.91*
Difficulty level						
	4.64 (1.12)	4.62 (1.10)	0.00	3.98 (0.97)	5.21 (0.87)	163.30***
Lecturer impression						
What is your overall impression of the lecturer?	4.90 (1.23)	4.75 (1.30)	1.67	**5.08 (1.15)**	**4.58 (1.32)**	16.90***
What is your impression that the students think of the lecturer?	4.69 (1.18)	4.62 (1.21)	0.54	**5.04 (1.01)**	**4.29 (1.25)**	40.64***
To what extent do you think the lecturer is engaged in the teaching?	**5.73 (1.30)**	**5.42 (1.48)**	4.50*	**5.72 (1.29)**	**5.42 (1.49)**	4.52*
Do you think the lecturer is a serious person?	5.57 (1.49)	5.46 (1.60)	0.25	4.73 (1.57)	6.22 (1.13)	112.37***
Do think the lecturer is knowledgeable?	6.07 (1.14)	6.07 (1.13)	0.05	5.67 (1.21)	6.44 (0.91)	47.94***
How does the lecturer seem to be as a leader of the teaching team?	3.92 (1.61)	3.91 (1.62)	0.00	3.71 (1.63)	4.10 (1.62)	5.21*
Competence	5.74 (1.15)	5.53 (1.37)	2.33	5.29 (1.31)	5.94 (1.16)	24.81***
Boring (scale 1–4)						
Preparation and organization	**3.01 (0.84)**	**2.82 (0.90)**	4.27*	2.73 (0.87)	3.07 (0.85)	13.66***
Quality of instructional material	3.23 (0.67)	3.09 (0.77)	3.14	3.06 (0.73)	3.24 (0.72)	5.08*
Ability to encourage work	2.85 (0.84)	2.76 (0.87)	0.50	**2.95 (0.81)**	**2.66 (0.88)**	11.10***
Availability	3.14 (0.87)	3.14 (0.92)	0.00	**3.38 (0.78)**	**2.92 (0.93)**	28.70***
Quality of contact	2.98 (0.88)	2.89 (0.89)	1.07	**3.29 (0.76)**	**2.60 (0.87)**	71.45***
Ability to lead the class	2.72 (0.90)	2.59 (0.89)	1.77	2.42 (0.88)	2.87 (0.85)	24.11***
Ability to relate to current issues	**2.80 (0.85)**	**2.57 (0.86)**	6.99**	**2.84 (0.76)**	**2.54 (0.93)**	12.04***
Contribution to the students' intellectual development	**3.02 (0.81)**	**2.83 (0.81)**	5.35*	2.97 (0.73)	2.87 (0.89)	1.95
Single items						
Overall impression of the course	**5.01 (1.15)**	**4.67 (1.26)**	7.68**	4.93 (1.16)	4.74 (1.26)	2.89
Would you like to attend a course with the lecturer	4.53 (1.57)	4.46 (1.71)	0.31	**4.83 (1.49)**	**4.19 (1.72)**	15.13***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Bold figures indicate significant differences in favor of a woman/feminine lecturer, gray highlighting indicate significant differences in favor of a masculine lecturer.

availability, both of which comply to a nursing, care-taking feminine gender role (Abele and Wojciszke, 2014). Finally, the masculine lecturer received higher ratings on organization and preparation.

It should, however, be noted that the feminine and masculine descriptions do not describe gender *per se*, but rather traits and behaviors associated with gender. This is interesting, because the behavior seemed to be more important than the lecturer's gender, and also more important than whether a lecturer engages in congruent or incongruent gender behavior. In short, behavior and characteristics seem to trump gender information regarding how the lecturers in our study were evaluated, however, the evaluations still follow stereotypical patterns of femininity and masculinity. Moreover, gender information and gender stereotypical behavior and characteristics sometimes seem to clash, potentially leading to a very precarious situation for lecturers in general.

DISCUSSION

Two experiments tested if the conflict between the gender role for women and the role of a university lecturer would be the reason that previous research has shown a general gender bias in SET. Previous research shows that women often receive lower SET compared to men, but also that SET follow gendered expectations (MacNell et al., 2015; Boring et al., 2016; Mengel et al., 2018; Mitchell and Martin, 2018). This article makes several important contributions. First, we use an experiment manipulating gender congruency in behavior, second, even though our hypotheses were not supported, the results highlight new knowledge about the gendered nature of SET, and thereby also contributes to the on-going discussion about SET and their use. In two experiments, we found that evalutions of a target lecturer depended on their stereotypically gendered displayed behavior and described characteristics, and that these evaluations heavily followed gendered expectations.

Much research in social psychology shows that women and men are thought to possess different traits and characteristics that correspond to general behaviors displayed by their respective gender group on an aggregated level (Ellemers, 2018). When there is a lack of fit or incongruence between the stereotypical ideas of how someone should be or behave, in regards to gender, and the stereotypical associations to the role they hold, this incongruence may lead to biases and criticism (Heilman, 1983, 2001, 2012). The lack of fit can be driven by actual job segregation (such as in this case, where more men than women are observed

in the role of university lecturers) or stereotypical ideas that a university lecturer is a man, as we found in the pilot study. Hence, we expected that women lecturers overall would receive lower SET than men, because a lack of fit between gender stereotypes and professional role. Second, we hypothesized that a woman lecturer described as masculine in terms of behavior and characteristics would be rated lowest on SET, because of the major violation of gender norms. However, none of the hypotheses were supported.

Hence it seems that in this situation, violations of gender roles and behavior does not seem to elicit negative perceptions of the lecturer. This points to a positive development within the context of higher education since it implies that both women and med can engage in both gender stereotypical and non-stereotypical behavior without being punished (or rewarded) through SET. This means that from this study we can not say that it is an inconsistency between women lecturers' behavior that has led to the generally lower SET for women that has previously been observed (MacNell et al., 2015; Boring et al., 2016; Mengel et al., 2018). We suggest that more studies should be performed to truly establish that this is the case.

There was a fairly consistent and strong pattern that the described behavior and characteristics influenced evaluations, although not in the hypothesized direction. Instead, the feminine behavior was at large evaluated more positively, compared to the masculine behavior. Nonetheless, the pattern makes sense from a gender stereotype perspective. Overall, the ratings conformed to gender stereotypes about femininity and masculinity, even though there were some differences between the experiments. In Experiment 1, the feminine condition led to better, more positive evaluations almost across the board of questions. However, higher work load, demands and requirements were more strongly associated with the masculine behavior. These are not necessarily indicative of negativity, but are more clearly associated with a masculine stereotype of being stern, assertive, and demanding (Abele and Wojciszke, 2014). Still, the participants strongly preferred the lecturer with feminine behavior, despite the lecturer's gender. As mentioned, one reason for the overwhelmingly positive evaluations of the feminine behavior in Experiment 1, could be the assymmetric description with respect to valence where the feminine version did not include many negative aspects, while the masculine version included few positive aspects, at least from a student perspective. For instance, in the masculine condition, the lecturer was presented as a leading researcher, which is not necessarily something that the students care about. Hence, the results of Experiment 1 should be interpreted with caution.

Nevertheless, the tendencies identified in Experiment 1 were at large confirmed in Experiment 2, where the stimuli material was more ambiguous in terms of valence. Because stereotypes are heuristics in impression formation (Heilman, 2012), evaluators may rely more heavily on them when there is little or ambiguous information. The results of the second experiment were accordingly slightly different, but the general pattern showed that evaluations largely conformed to gender stereotypes. The lecturer described as masculine was percieved as a better leader, more competent, a better pedagogue, "tougher," and students expected to learn more from their course. Hence, evaluations of the masculine behavior followed mainly from stereotypically masculine attributes such as leadership skills, competence and goal-orientation (Abele and Wojciszke, 2014). However, the feminine lecturer was percieved as being more approachable and was more liked. Moreover, and similar to the Experiment 1, participants preferred to attend the course when the lecturer was a woman. Again, these features conform to a feminine gender stereotype which is focused on the maintenance of relationships (Abele and Wojciszke, 2014).

These two experiments highlight the precarious situation that lecturers may face. While the feminine behavior increased liking, the masculine behavior increased competence ratings. Even though there were no interactions with the lecturer's gender, it is plausible to assume that this balance is more difficult for women lecturers where the likable traits and behaviors are expected, and cannot be bargained with (Heilman and Okimoto, 2007). It may be difficult for a lecturer to be rated good on both liking (or warmth) and competence, which is in line with research on gender stereotypes (Fiske et al., 2007; Heilman, 2012). Given that SET form the basis of hiring and promotion decisions (Cashin, 1999; Seldin, 1999; Clayson, 2009; Davis, 2009; Seldin et al., 2010), the results of the present reseach contributes to the literature.

Much of the international research on SET use questions specifically about lecturers as individuals, and their traits (Goebel and Cashen, 1979; Ryan and Wilson, 2014; Wilson et al., 2014; MacNell et al., 2015). However, whether a person is seen as compassionate or caring does not reveal information about their ability to perform as a lecturer, or about their pedagogical skills, which should be the focus of SET, regardless of how SET are to be used. Therefore, other questions should be given space, such as questions relating to the set-up of the course, the organization, the study materials etc. It is plausible to believe that such evaluations would better estimate a lecturer's pedagogical skills and abilities. However, as shown in the two experiments in this article, these judgements still obey to gendered expectations about behavior. These results line up with previous research by Boring et al. (2016) and Mitchell and Martin (2018) who found that a gender bias affected judgment of seemingly objective aspects of teaching.

Limitations and Suggestions for the Future

To our knowledge, this is the first experimental design that test gender bias in SET. The benefit of using experiments in research is also their drawback—the setting is sterile and context-free. The positive side is that the experiment allows for high control over potential confounds. In this first attempt, we aimed to have as little confounding information as possible. Hence, the stimuli material did not, for example, present what field the lecturer is active in, which is a factor previous shown to affect gender bias in SET (Boring et al., 2016; Mengel et al., 2018). This implies that the description may be too "clean" and generic, which might result in difficulties for the participants to truly engage in the described lecturer. Because the lack of substantial information to relate the lecturer to, this may lead to social desirability—that answers are colored by a desire to appear gender egalitarian. In line with this, the expected effects of the lecturer's gender were not found in any of the experiments, nor were the interactions between gender and incongruent behavior. One reason may be

that the participants were aware of gender aspects in these kinds of situations, which could lead to socially desireable answers. Indicative of this interpretation is that when the participants were asked to indicate their thoughts regarding the purpose of the study, several suggested that the study regarded gender issues. Hence, we also suspect that the gender manipulation may be more strongly influenced by social desireability compared to the behavior manipulation. Future studies may apply a more subtle way to manipulate the lecturer's gender, perhaps by using a photo of the lecturer.

There were some inconsistencies between the results found in Experiment 1 and 2, which probably were due to the non-balanced valence of the descriptions used in Experiment 1. From a student perspective, a lecturer who is engaged with the teaching, being caring and responsive should lead to higher ratings. Therefore, the results from Experiment 2 is more informative. It would be beneficial to develop the descriptions more, and for instance describe a lecturer as having both feminine and masculine behaviors. We believe this to be important knowledge for all researchers conducting this kind of text-based experiments.

Conclusions

The present study showed that behavior and characteristics seem to trump the lecturer's gender in SET, at least in this kind of relatively artificial experimental setting. This result could be interpreted as a positive outcome, since evaluations are based on behavior, rather than gender of the lecturer. Nonetheless, the evaluations of behavior follow gender stereotypes, where a lecturer described as showing masculine behavior was also seen as possessing characteristics such as competence and professionalism, whereas a lecturer described as showing feminine behavior also was seen as possessing characteristics such as being caring and nurturing. In this way, the results of this research align with social psychological theory on gender stereotypes (Eagly and Wood, 2012; Abele and Wojciszke, 2014).

However, the participants displayed somewhat contradictory responses in that they liked the caring and nurturing (i.e., the feminine) lecturer better, although they gave the masculine lecturer higher ratings on work performance. This finding is problematic, because it leaves the individual lecturer in a difficult situation. Should a lecturer focus on being professional and making sure that students actually learn, or should they be accommodating and responsive, which hence results in being liked and increases students' desire for attending the course. Therefore, these kinds of results should be communicated not only to lecturers, but also to students, so they can be aware of their own biases. The finding contributes to the ongoing discussion about the validity of SET in judging individual lecturers' pedagogical skills (Yunker and Yunker, 2003; Wright, 2006; Beecham, 2009; Hoefer et al., 2012; Spooren et al., 2013; Braga et al., 2014; Stark and Freishtat, 2014; Boring et al., 2016). Given the results of the present study, there is an urge to develop reliable and valid measures of SET. To some extent, the ratings seem to fall out on two dimensions, where for instance the lecturer's availability and ability to encourage may not necessarily go along with their pedagogical skills, such as course set-up, materials, leadership etc. We therefore join the scholars before us, and raise critical voices regarding the use of SET in their current form as the main tool for assessing lecturers' pedagogical skills and abilities, for instance regarding hiring or promotion purposes. If SET are to be used for such purposes, they should be further developed and validated to better capture actual ability of a lecturer and not reflect popularity or biases. For instance, collegial evaluations, exam results, or performance in subsequent courses could be used to validate SET, and comprise part of the evaluation of a lecturer's competence.

However, it is important to remember that SET were introduced for formative purposes, that is, to improve the teaching and student-relations (Hornstein, 2016). In that sense, SET may be better used. It is important that teachers and students share a common goal in the teaching process and that the student perspective is present when courses are developed.

We believe that two main important outcomes of this article should be highlighted. First, this is to our knowledge the first attempt to make causal inferences regarding the mechanism behind gender biases in SET, using a strict experimental paradigm. Second, we find that gender information does not seem to evoke negative evaluations of women lecturers on a general level. Moreover, gender incongruent behavior is not sanctioned by lower SET. However, students' ratings are somewhat contradictory in that they prefer a lecturer that they see as less competent and pedagogically skilled. This could leave individual lecturers in a difficult position.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found at the Open Science Framework: https://osf.io/sfcym/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ER contributed to the general idea, first design, analyses, and manuscript drafts. MG and AL contributed to finalizing the design, continuous discussions about methods and results, and finalizing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by a grant from the Swedish Research Council for Work Life and Welfare, Grant No. 253099-131526.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2020.571287/full#supplementary-material

REFERENCES

Abele, A. E., and Wojciszke, B. (2014). Communal and agentic content in social cognition: a dual perspective model. *Adv. Exp. Psychol.* 50, 195–255. doi: 10.1016/B978-0-12-800284-1.00004-7

Adams, J. V. (1997). Student evaluations: the ratings game. *Inquiry* 1, 10–16.

Andersen, K., and Miller, E. D. (1997). Gender and student evaluations of teaching. *Polit. Sci. Polit.* 30, 216–219. doi: 10.1017/S1049096500043407

Bachen, C. M., McLoughlin, M. M., and Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Commun. Educ.* 48, 193–210. doi: 10.1080/03634529909379169

Bakan, D. (1966). *The Duality of Human Existence. An Essay on Psychology and Religion*. Chicago, IL: Rand Mcnally.

Bartol, K. M., and Butterfield, D. A. (1976). Sex effects in evaluating leaders. *J. Appl. Psychol.* 61, 446–454. doi: 10.1037/0021-9010.61.4.446

Beecham, R. (2009). Teaching quality and student satisfaction: nexus or simulacrum? *Lond. Rev. Educ.* 7, 135–146. doi: 10.1080/14748460902990336

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *Int. J. Teach. Learn. High. Educ.* 17, 48–62.

Boring, A. (2016). Gender biases in student evaluations of teaching. *J. Public Econ.* 145, 27–41. doi: 10.1016/j.jpubeco.2016.11.006

Boring, A., Ottoboni, K., and Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Res.* 1–11. doi: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Braga, M., Paccagnella, M., and Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Econ. Educ. Rev.* 41, 71–88. doi: 10.1016/j.econedurev.2014.04.002

Brescoll, V., Dawson, E., and Uhlmann, E. L. (2010). Hard won and easily lost: the fragile status of leaders in gender-stereotype-incongruent occupations. *Psychol. Sci.* 21, 1640–1642. doi: 10.1177/0956797610384744

Carli, L. L. (1990). Gender, language and influence. *J. Pers. Soc. Psychol.* 59, 941–951. doi: 10.1037/0022-3514.59.5.941

Carli, L. L., LaFleur, S. J., and Loeber, C. C. (1995). Nonverbal behavior, gender, and influence. *J. Pers. Soc. Psychol.* 68, 1030–1041. doi: 10.1037/0022-3514.68.6.1030

Cashin, W. E. (1999). "Student ratings of teaching: uses and misuses," in Current Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions, ed P. Seldin (Bolton, MA: Anker), 25–44.

Centra, J. A., and Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *J. Higher Educ.* 71, 17–33. doi: 10.1080/00221546.2000.11780814

Chamberlin, M. S., and Hickey, J. S. (2001). Student evaluations of faculty performance: the role of gender expectations in differential evaluations. *Educ. Res. Q.* 25, 3–14.

Clayson, D. E. (2009). Student evaluations of teaching: are they related to what students learn? A meta-analysis and review of the literature. *J. Market. Educ.* 31, 16–30. doi: 10.1177/0273475308324086

Davis, B. G. (2009). *Tools for Teaching, 2nd Edn*. San Francisco, CA: John Wiley & Sons.

Eagly, A. H., and Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychol. Rev.* 109, 573–598. doi: 10.1037/0033-295X.109.3.573

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., and Sczesny, S. (2020). Gender stereotypes have changed: a cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *Am. Psychol.* 75, 301–315. doi: 10.1037/amp0000494

Eagly, A. H., and Wood, W. (2012). "Social role theory," in *Handbook of Theories of Social Psychology*, eds P. A. M. van Lange, A. W. Kruglanski, and E. T. Higgins (London: Sage), 458–476.

Ellemers, N. (2018). Gender stereotypes. *Annu. Rev. Psychol.* 69, 275–298. doi: 10.1146/annurev-psych-122216-011719

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., et al. (2019). Gender and cultural bias in student evaluations: why representation matters. *PLoS ONE* 14:e0209749. doi: 10.1371/journal.pone.0209749

Feldman, K. A. (1993). College students' views of male and female college teachers: evidence from the social laboratory and experiments – Part 2. *Res. High. Educ.* 34, 151–211. doi: 10.1007/BF00992161

Fiske, S. T., Cuddy, A. J. C., and Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* 11, 77–83. doi: 10.1016/j.tics.2006.11.005

Galbraith, C., Merrill, G., and Kline, D. (2012). Are student evaluations of teaching effectiveness valid for measuring student outcomes in business related classes? A neural network and Bayesian analyses. *Res. Higher Educ.* 53, 353–374. doi: 10.1007/s11162-011-9229-0

Goebel, B. L., and Cashen, V. M. (1979). Age, sex, and attractiveness as factors in student ratings of teachers: a developmental study. *J. Educ. Psychol.* 71, 646–653. doi: 10.1037/0022-0663.71.5.646

Gustafsson Sendén, M., Klysing, A., Lindqvist, A., and Renström, E. A. (2019). The (not so) changing man: dynamic gender stereotypes in Sweden. *Front. Psychol.* 10:37. doi: 10.3389/fpsyg.2019.00037

Heilman, M. E. (1983). Sex bias in work settings: the lack of fit model. *Res. Organ. Behav.* 5, 269–298.

Heilman, M. E. (2001). Description and prescription: how gender stereotypes prevent women's ascent up the organizational ladder. *J. Soc. Issues* 57, 657–674. doi: 10.1111/0022-4537.00234

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Res. Organ. Behav.* 32, 113–135. doi: 10.1016/j.riob.2012.11.003

Heilman, M. E., and Chen, J. J. (2005). Same behavior, different consequences: reactions to men's and women's altruistic citizenship behavior. *J. Appl. Psychol.* 90, 431–441. doi: 10.1037/0021-9010.90.3.431

Heilman, M. E., and Haynes, M. C. (2005). No credit where credit is due: attributional rationalization of women's success in male-female teams. *J. Appl. Psychol.* 90, 905–916. doi: 10.1037/0021-9010.90.5.905

Heilman, M. E., and Okimoto, T. G. (2007). Why are women penalized for success at male tasks?: the implied communality deficit. *J. Appl. Psychol.* 92, 81–92. doi: 10.1037/0021-9010.92.1.81

Hentschel, T., Heilman, M. E., and Peus, C. V. (2019). The multiple dimensions of gender stereotypes: a current look at men's and women's characterizations of others and themselves. *Front. Psychol.* 10:11. doi: 10.3389/fpsyg.2019.00011

Hoefer, P., Yurkiewicz, J., and Byrne, J. C. (2012). The association between students' evaluation of teaching and grades. *Decis. Sci. J. Innov. Educ.* 10, 447–459. doi: 10.1111/j.1540-4609.2012.00345.x

Hornstein, H. A. (2016). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Educ.* 4:1304016. doi: 10.1080/2331186X.2017.1304016

Jago, A. G., and Vroom, V. H. (1982). Sex differences in the incidence and evaluation of participative leader behavior. *J. Appl. Psychol.* 67, 776–783. doi: 10.1037/0021-9010.67.6.776

Lindqvist, A., Gustafsson Sendén, M., and Renström, E. A. (2020). What is gender, anyway: a review of the options for operationalising gender. *Psychol. Sex.* doi: 10.1080/19419899.2020.1729844. [Epub ahead of print].

MacNell, L., Driscoll, A., and Hunt, A. N. (2015). What's in a name: exposing gender bias in student ratings of teaching. *Innov. Higher Educ.* 40, 291–303. doi: 10.1007/s10755-014-9313-4

Makarova, E., Aeschlimann, B., and Herzog, W. (2019). The gender gap in STEM fields: the impact of the gender stereotype of math and science on secondary students' career aspirations. *Front. Educ.* 4:60. doi: 10.3389/feduc.2019.00060

Mengel, F., Sauermann, J., and Zolitz, U. (2018). Gender bias in teaching evaluations. *J. Eur. Econ. Assoc.* 17, 535–566. doi: 10.1093/jeea/jvx057

Mitchell, K. M. W., and Martin, J. (2018). Gender bias in student evaluations. *Polit. Sci. Polit.* 51, 648–652. doi: 10.1017/S104909651800001X

Rowden, G. V., and Carlson, R. E. (1996). Gender issues and students' perceptions of instructors' immediacy and evaluation of teaching and course. *Psychol. Rep.* 78, 835–839. doi: 10.2466/pr0.1996.78.3.835

Rudman, L. A. (1998). Self-promotion as a risk factor for women: the costs and benefits of counterstereotypical impression management. *J. Pers. Soc. Psychol.* 74, 629–645. doi: 10.1037/0022-3514.74.3.629

Rudman, L. A., and Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *J. Soc. Issues* 57, 743–762. doi: 10.1111/0022-4537.00239

Rudman, L. A., Moss-Racusin, C. A., Phelan, J. E., and Nauts, S. (2012). Status incongruity and backlash effects: defending the gender hierarchy motivates prejudice against female leaders. *J. Exp. Soc. Psychol.* 48, 165–179. doi: 10.1016/j.jesp.2011.10.008

Ryan, R., and Wilson, J. H. (2014). Professor-student rapport scale: psychometric properties *of* the brief version. *J. Scholarship Teach. Learn.* 14, 64–74. doi: 10.14434/josotl.v14i3.5162

Seldin, P. (1999). "Building successful teaching evaluation programs," in *Changing Practices in Evaluating Teaching: A Practical Guide to Improved Faculty*

Performance and Promotion/Tenure Decisions, ed P. Seldin (Boston, MA: Anker), 213–242.

Seldin, P., Miller, J. E., and Seldin, C. A. (2010). *The Teaching Portfolio: A Practical Guide to Improved Performance and Promotion/ Tenure Decisions, 4th Edn*. San Francisco, CA: Jossey-Bass.

Spooren, P., Brockx, B., and Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Rev. Educ. Res.* 83, 598–642. doi: 10.3102/0034654313496870

Sprague, J., and Massoni, K. (2005). Student evaluations and gendered expectations: what we can't count can hurt us. *Sex Roles* 53, 779–793. doi: 10.1007/s11199-005-8292-4

Stark, P. B., and Freishtat, R. (2014). *An Evaluation of Course evaluations. ScienceOpen Res*. 1–7. doi: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Wilson, J. H., Beyer, D., and Monteiro, H. (2014). Professor age affects student ratings: haloeffect for younger teachers. *Coll. Teach.* 62, 20–24. doi: 10.1080/87567555.2013.825574

Wright, R. E. (2006). Student evaluations of faculty: concerns raised in the literature, and possible solutions. *Coll. Stud. J.* 40, 417–422.

Yunker, P. J., and Yunker, J. A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *J. Educ. Bus.* 78, 313–317. doi: 10.1080/08832320309598619

Zhen, W., Kark, R., and Meister, A. L. (2018). Paradox versus dilemma mindset: a theory of how women leaders navigate the tensions between agency and communion. *Leadership Q.* 29, 584–596. doi: 10.1016/j.leaqua.2018.04.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.