This is the peer reviewed version of an article that has been accepted for publication in Sex Roles. An uncorrected and unedited version, it is not the version of reference.

Gender Bias in Student Evaluations of Teaching: Students' Self-Affirmation Reduces the Bias by Lowering Evaluations of Male Professors

Vera Hoorens, Gijs Dekkers, and Eliane Deschrijver

KU Leuven

Author Note

Vera Hoorens (ORCID 0000-0002-4855-9861), Center for Social and Cultural Psychology, KU Leuven; Gijs Dekkers (ORCID 0000-0003-0566-8723), Center for Sociological Research, KU Leuven; Eliane Deschrijver (ORCID 0000-0003-0387-3539), Center for Social and Cultural Psychology, KU Leuven

Eliane Deschrijver is now at Ghent University.

This work was supported by the Special Research Fund of KU Leuven [OT/12/38], awarded to the first author. The research has been presented at the 18th Annual Conference of the Society for Personality and Social Psychology, San Antonio, Texas, 19-21 January 2017 as part of the Symposium *The dark side of striving for and experiencing uniqueness*, organized by the first author. The authors thank the following individuals for their help in different stages of the study: Maarten Borgers, Liedewij Borremans, Toon Bynens, Joke Claes, Quentin Clemens, Ruben Habex, Carolien Van Damme, and Nathalie Vissers.

Correspondence concerning this manuscript should be addressed to Vera Hoorens, KU Leuven, Center for Social and Cultural Psychology, Tiensestraat 102, mailbox 3727, B-3000 Leuven, Belgium. Email: Vera.Hoorens@kuleuven.be

Abstract

Students evaluate male professors higher than female professors. In a study that we presented to participants as a test of a new form for student evaluations of teaching (SETs), we examined if self-affirmation (contemplating elements that positively contribute to one's self-image) reduced the gender bias. Belgian students (n = 568), who were randomly assigned to self-affirm (through either a value-affirmation task or self-superiority priming) or not, read a vignette prompting them to imagine that they had received a good or a bad grade from a male or a female professor. They evaluated the course, the professor, and the form. Non-self-affirmed participants showed a gender bias after a bad grade, disadvantaging the female professor. Self-affirmation eradicated the gender bias by lowering evaluations for the male professor, suggesting that the gender bias involves overvaluing male rather than derogating female professors. Without self-affirmation, the positivity of the SETs was correlated with participants' evaluation of the SET form itself. Self-affirmation inflated the correlation for the male professor and eradicated it for the female professor. Having students self-affirm before SETs may be useful when SETs are obligatory only. An even better approach is asking SETs before students learn their grades or simply abolish SETs as a factor in hiring and promotion decisions.

Keywords: sexism; prejudice; course evaluation; teacher effectiveness evaluation; students' evaluation of teaching; self-affirmation

Gender Bias in Student Evaluations of Teaching: Students' Self-Affirmation Reduces the Bias by Lowering Evaluations of Male Professors

Universities and colleges around the world give students the opportunity to evaluate their professors' teaching. These student evaluations of teaching (SETs) are highly consequential for individual professors because they play a major role in decisions concerning hiring, tenure, and promotion (Stroebe, 2016). It is therefore critical that they allow a valid and fair assessment of the quality of teaching. However, SETs are often largely unrelated to, and sometimes even negatively correlated with, more objective measures of teacher effectiveness such as grades obtained in follow-up courses (Boring, Ottoboni, & Stark, 2016; Braga, Paccagnella, & Pellizzari, 2014; Carrell & West, 2010).

The problematic validity of SETs is partly due to their sensitivity to factors that are irrelevant for quality of teaching (Boring et al., 2016). One well-documented bias has to do with the professor's gender, with male professors receiving higher scores on SETs than female professors (Abel & Meltzer, 2007; Arbuckle & Williams, 2003; Basow & Silberg, 1987; Boring, 2017; Fisher, Stinson, & Kalajdzic, 2019; MacNell, Driscoll, & Hunt, 2015; Mengel, Sauermann, & Zölitz, 2019; Nadler, Berry, & Stockdale, 2013; Pounder, 2007; Sidanius & Crane, 1989; Wagner, Rieger, & Voorvelt, 2016). For schools and universities that value well-informed and just decisions about faculty, it is important to develop strategies to avoid or reduce this gender bias and other biases that threaten the validity of SETs. We tested the effect of two self-affirmation strategies, one a well-established value-affirmation task and the other a novel self-superiority priming procedure. We did so in a vignette study wherein we adapted an experimental design from Sinclair and Kunda (2000).

(Gender) Bias in SETs

The landscape of factors that bias SETs includes seemingly trivial characteristics of the circumstances in which students provide them. For example, good weather on the day of the evaluation entails higher SETs (Braga et al., 2014). In one study, students who had received a chocolate bar (from someone external to the course) gave higher SETs than students who had not received such a treat (Youmans & Jee, 2007).

Whereas circumstantial factors do not necessarily entail unfair SETs, professor-related factors do. Arguably the best known factor in this category is grading leniency. Professors who give higher grades typically obtain higher SETs (Carrell & West, 2016; Krautmann & Sander, 1999; Weinberg, Hashimoto, & Fleisher, 2009; Wright & Jenkins-Guarnieri, 2012). Grading is often thought to be under professors' control, but in many cases it is not completely so. At European universities, for example, many bachelor programs do not have strict admission criteria. First-year courses are thus populated by students with widely varying academic abilities, attitudes, and levels of motivation. A considerable number fail courses or obtain grades that barely allow them to pass. On average, therefore, the grades awarded by professors teaching introductory courses are lower than those awarded by professors teaching advanced courses (after the weakest students have dropped out) or teaching in more selective programs. Thus, the proportion of grading leniency that professors have under their control is more limited than it might seem.

A particularly problematic category of determinants includes personal characteristics that completely escape professors' control. Among them are the professor's age and ethnicity. Students tend to give higher SETs to younger than to older professors (Arbuckle & Williams, 2003; Joye & Wilson, 2015), and to White professors than to Professors of Color or from ethnic

minorities (Fan et al., 2019; Reid, 2010). However, perhaps the most widely documented factor is the professor's gender.

Students generally evaluate male professors more favorably than female professors (Basow & Silberg, 1987; Fisher et al., 2019; MacNell et al., 2015; Sidanius & Crane, 1989; Wagner et al., 2016). Some researchers have found that this gender bias is mostly driven by SETs given by male students (Boring, 2017; Mengel et al., 2019), but other researchers found that both female and male students favor male professors (Abel & Meltzer, 2007; Arbuckle & Williams, 2003; Nadler et al., 2013). Evidence for the idea that the gender bias in SETs is rooted in a gendered stereotype of the "good professor" comes from research where students' expectations about professors' teaching were examined. In general, students expected to learn more from male than from female professors (Clayson, 2019). Still, the gender bias remains controversial. In some studies, SETs of female professors were similar to, or higher than, those of male professors (Bachen, McLoughlin, & Garcia, 1999; Bavishi, Madera, & Hebl, 2010; Centra & Gaubatz, 2000). Some authors have therefore concluded that professors' gender does not affect SETs (Feldman, 1992, 1993).

One explanation for the conflicting findings is that actual gender differences in teaching practices, academic positions, disciplines, and subject matters may obscure the gender bias in some studies and inflate it in other ones. Some researchers have therefore controlled for actual differences by studying SETs across disciplines and academic ranks (Boring, 2017) or by including matched samples of male and female professors (Fisher et al., 2019). Others have asked students to evaluate professors whom they had never met (e.g., professors teaching online courses; MacNell et al., 2015), whom they had only heard on an audio tape (Arbuckle & Williams, 2003), or of whom they had only read a written lecture (Abel & Meltzer, 2007).

Studies using any of those strategies consistently yield a gender bias (Abel & Meltzer, 2007; Arbuckle & Williams, 2003; Boring, 2017; Fisher et al., 2019; MacNell et al., 2015).

Another explanation is that different studies have used different measures, ranging from general impressions of teaching quality to judgments of specific behaviors and attitudes (e.g., showing a personal interest in students; Bachen et al., 1999). General impressions are more abstract and less verifiable than specific judgments (Fiedler & Semin, 1988) and thus allow a greater latitude for bias. Studies that have yielded strong gender biases in SETs have indeed used judgments of general impressions, sometimes using just one item (Peterson, Biederman, Andersen, Ditonto, & Roe, 2019; Sinclair & Kunda, 2000, Study 1; Wagner et al., 2016).

A particularly interesting explanation for the conflicting findings is that gender bias in SETs may above all manifest itself when it can serve to bolster students' self-views. Research in other contexts has shown that people show more prejudice after interpersonal criticism has threatened their feelings of self-worth (Allen & Sherman, 2011; Collange, Benbouzyane, & Sanitioso, 2006; Collange, Fiske, & Sanitioso, 2009; Fein & Spencer, 1997). In all these studies, the source of the criticism differed from the target of the prejudice. The research of Sinclair and Kunda (2000, p. 1329) on gender bias in SETs was the first known to examine what they called "motivated stereotyping" targeting the individual from whom the self-threat emanated. Their reasoning was that bad grades threaten students' feelings of self-worth. Students can bolster their self-esteem by derogating the source of their self-threat, that is, by giving a negative SET to the professor who has awarded the grade. If the professor is a woman, the circumstance that she deviates from the stereotype of the good professor (who for many students continues to be a man) then comes in handy.

To test this hypothesis, Sinclair and Kunda (2000, Study 1) examined SETs for real-life courses, given shortly after the grades for a given term had been announced. The study replicated the grade effect in that students gave lower SETs to professors who had given them a bad grade than to professors who had given them a good grade. However, the grade effect was larger for female than for male professors. The students evaluated male and female professors similarly if they had received a good grade, but gave female professors lower SETs than male professors if they had received a bad grade. This gender bias could not be due to different course domains or difficulty levels because students with good and bad grades gave SETs for overlapping sets of courses. On average, moreover, female and male professors did not give different grades. Sinclair and Kunda (2000) thus convincingly demonstrated a gender bias among students whose feelings of self-worth had been threatened by a bad grade.

One question that research on the gender bias to date cannot answer is whether the gender bias involves a bias against female professors or a bias in favor of male professors. Sinclair and Kunda (2000) presented motivated prejudice as a derogation of female professors. However, it is also possible that SETs for male professors are inflated or that SETs for professors of both genders are biased in different directions. In the case of gender bias after a bad grade, the grade may in principle provoke derogation of any professor. However, the fact that the professor is a women does not necessarily add to that derogation. Instead, the derogation may be mitigated or even eradicated if the professor is man. Although the issue may seem purely academic—after all, the critical finding is that students evaluate male and female professors differently—it is potentially consequential. Efforts to reduce the bias may be more likely to be successful if they are informed by knowledge about its precise nature. If the gender bias is driven by derogation of female professors, efforts to eradicate it should arguably target evaluations of women; if it is

driven by approbation of male professors, they should arguably target evaluations of men (or work to accord women the same status protection afforded men).

Reducing Gender Bias

Some authors have suggested to statistically partial out biases from SET scores (Greenwald & Gillmore, 1997). Yet, there is no easy manner to do so for gender bias. In any given SET it is hard to determine the relative contributions of actual teaching quality and professors' gender. By consequence, it may be hard to make male professors accept that SETs of female professors are adjusted upward or that their own SETs are adjusted downward. It is therefore important to develop interventions that prevent (rather than post-hoc correct for) gender bias. One approach that has been tested is enriching the instructions for SETs with an explanation of "unconscious and unintentional" gender and race biases and the request to the students to resist stereotypes in their evaluations. In a field experiment, this approach reduced gender bias in real-life SETs by enhancing the evaluation of female professors (Peterson et al., 2019).

However, one potential side-effect of warning students against the gender bias (or other biases) is overcorrection. Overcorrection, a bias in the direction opposite to the distortion for which one tries to compensate, has been observed in various other contexts (Echterhoff, Groll, & Hirst, 2007; Petty, Wegener, White, 1998; Seta, Seta, & McCormick, 2020; Sommers & Kassin, 2001). Of direct relevance to the present context, it may occur when people become aware that their behavior toward members of certain groups may be prejudiced (Chien, Wegener, Petty, & Hsiao, 2014; Mendes & Koslov, 2013). In the case of SETs, overcorrection for the gender bias might entail unfairly high evaluations of female professors.

On the other hand, warning students against the gender bias may exacerbate the very gender bias it seeks to eradicate. One reason is that it may render thoughts about the professor's gender more salient (Crandall & Eshleman, 2003). The professor's gender may thus affect their evaluations more than it would otherwise do. Another reason is that a warning may be construed as an explicit request to evaluate generously if they are women and harshly if they are men. Seemingly needless, inappropriate, or coercive requests sometimes elicit reactance—the motivation to protect or regain one's personal freedom (Miron & Brehm, 2006; Mühlberger & Jonas, 2019). This motivation may elicit the tendency to do the opposite of what the external party seems to want. In one study, for example, participants expressed higher scores on racism measures after having read statements against racism than in a control condition (Legault, Gutsell, & Inzlicht, 2011). If students feel the target of inappropriate or coercive attempts to make them judge male professors harshly and/or female professors mildly, reactance may lead them to judge male professors mildly and/or female professors harshly.

A potentially useful strategy to avoid gender bias in SETs can be derived from the observation that it seems rooted in a male-gendered stereotype of professors being activated and applied in response to a self-threat (that is, receiving a bad grade). Research on coping with self-threats and research on motivated prejudice and stereotyping may therefore yield effective handles to reduce gender bias. One particularly promising phenomenon that has been identified in both fields is *self-affirmation*, broadly defined as engaging in an act that confirms one's self-integrity (i.e., an image of the self as moral and sufficiently competent to control the outcomes of important life experiences) (Cohen & Sherman, 2014).

The most widely used procedure to induce self-affirmation in research is a value-affirmation task (Cohen & Sherman, 2014; Sherman & Cohen, 2006). In that task, people review

their value-hierarchy (by rank ordering values) and contemplate how one of their top values affects their personal lives (by writing a paragraph or answering questions about it). Yet, several other self-affirmation procedures exist. Among them are having people think about their qualities (Cohen, Aronson, & Steele, 2000; Study 2; Harvey & Oswald, 2000) or judge the self on a series of character strengths (Napper, Harris, & Epton, 2009).

Self-affirmation has been shown to reduce stress and defensive reactions in the face of criticism (Cohen & Sherman, 2014; Critcher & Dunning, 2015; Epton, Harris, Kane, Koningsbruggen, & Sheeran, 2015; Sherman & Hartson, 2011). It has also been shown to make people more willing to apologize to someone they have wronged (Schumann, 2014) and less likely to respond aggressively when confronted by stigmatized individuals (Stone, Whitehead, Schmader, & Focella, 2011). Of particular importance here, self-affirmation reduces the expression of prejudice and intergroup bias (Badea & Sherman, 2019; Fein & Spencer 1997; Lehmiller, Law, & Tormala, 2010).

The Present Research

We set out to examine if self-affirmation reduces gender bias in SETs. To that end, we wished to experimentally manipulate the task preceding a SET. We designed a vignette study where students imagined that they had taken a course taught by a given professor and had obtained a certain grade. Besides avoiding ethical and legal complications of using real-life SETs, this approach gave us full control over the information that students received about the course and the professor. The information was identical in all conditions except that we manipulated grade and professors' gender. Half the participants read that they had obtained a good grade (good-grade condition), whereas the other half read that they had obtained a bad grade (bad-grade condition). In addition, half the participants read about a course taught by a

female professor (female-professor condition) whereas the other half read about a course taught by a male professor (male-professor condition). Any difference in evaluations of the professors within the grade conditions could thus be unequivocally attributed to gender stereotypes (MacNell et al., 2015).

We felt confident that a vignette study would yield valid data because earlier research has shown that the observation of a gender bias in SETs does not depend on methodology. Gender bias has been found in real-life SETs (Wagner et al., 2016), evaluations on websites such as RateMyProfessors.com (Fisher et al., 2019), and judgments of professors whom students had heard only once and had never seen (Arbuckle & Williams, 2003), professors of online courses (MacNell et al., 2015), and hypothetical professors (Kierstead, D'Agostino, & Dill, 1988). However, we established the validity of our vignette procedure by conceptually replicating the full design of Sinclair and Kunda (2000, Study 1). Although we were specifically interested in the responses of students imagining bad grades (bad-grade condition), we thus also included conditions where students imagined that they had obtained a good grade (good-grade condition). Also following Sinclair and Kunda (2000, Study 1), our dependent variables were general impressions of the course and the professor.

The cover story under which we introduced the study to participants ("a first exploration of the usability of a novel form that might be used for future SETs") required that we included many more questions in addition to our dependent variables (i.e., general evaluations of the course and the professor). We therefore included questions asking for more specific judgments of the course and the professor as well. Allegedly testing a novel evaluation form, we could not include items that often occur in SETs (e.g., about the adequacy of the course level). We therefore included questions about the professor's personality traits and the course's

characteristics. Our selection of questions was guided by the aim to lend the form some face validity in the eyes of the participants, but we did not expect any effects on them because their specificity greatly limited the latitude for any bias to occur. To further bolster the cover story, we also added questions about the quality of the form itself.

We compared responses to the vignette in a condition where participants had not done any prior task (baseline condition) with responses in conditions where participants had first gone through a self-affirmation procedure. We elicited self-affirmation through a classic value-affirmation task (value-affirmation condition), which we used because it is arguably the most widely used self-affirmation procedure (Cohen & Sherman, 2014), and through a novel self-superiority priming procedure (self-superiority condition). In the following we first explain this novel procedure itself before explaining why we decided to include it in our design.

The self-superiority priming procedure was inspired by the finding that most people believe that they are in many respects superior to others. Self-superiority beliefs take the form of thinking that one has a better personality (Alicke, 1985; Brown, 1986), acts more safely and healthily (Hoorens & Harris, 1998), and heads toward a rosier future than other people (Hoorens, Smits, & Shepperd, 2008; Weinstein 1980). It also takes the form of preferring one's attributes, even for the letters of one's name over other letters (Hoorens, 1990, 2014; Nuttin, 1985, 1987). Our self-superiority priming procedure involved having participants fill out a personality questionnaire with items that were selected to elicit particularly strong self-superiority responses. The difference between this procedure and earlier ones based on self-reported qualities is that we had participants judge their qualities (rather than taking them for granted and elaborate on them, as in Cohen et al., 2000, Study 2) and do so comparatively (rather than judging their qualities without the instruction to use social comparison, as in Napper et al., 2009). Self-superiority

beliefs have already been found to enhance feelings of self-worth (Brown, 1986, 2012). Contemplating and expressing one's self-superiority while filling out our questionnaire may be self-affirming in a similar manner as going through a value-affirmation procedure is.

For the present purposes (i.e., seeking to reduce the gender bias in SETs), the self-superiority procedure has three advantages as compared to the value-affirmation procedure. First, it arguably possesses great face validity. There is nothing conspicuous about having to do a self-evaluation (an evaluation that might in actual SETs be introduced as a self-assessment of one's personal development at that point in time) before evaluating an activity that presumably has contributed to one's personal development. Second, it can be applied multiple times in the course of an academic program. People arguably consider their qualities more changeable than their core values, such that the self-superiority procedure can be repeated without quickly seeming boring or needlessly repetitive. Third, the self-superiority questionnaire is brief and easy to fill out. The procedure thus requires little time and effort, which may enhance students' willingness to go through it as compared to more time-consuming self-affirmation procedures.

We tested two hypotheses. Hypothesis 1 stated that in the baseline condition, the difference between SETs given by students with a bad grade and by students with a good grade (lower SETs after a bad grade than after a good grade) would be larger if the professor was a woman than if the professor was a man. More specifically, it predicted a grade by professor gender interaction in the baseline condition such that in the good-grade condition, SETs will not differ between the female and the male professor and in the bad-grade condition, SETs will be higher if the professor is a man than if the professor is a woman.

Hypothesis 2 stated that value affirmation and self-superiority priming would eradicate the higher SETs for the male professor that, at baseline, occurred in the bad-grade condition but

not in the good-grade condition. Hypothesis 2 thus predicted a three-way interaction of initial task, grade, and professor gender such that (a) in the good-grade condition, SETs will not depend on initial task nor on professor gender and (b) in the bad-grade condition, SETs will jointly depend on initial task and professor gender. Participants will show a gender bias at baseline but not after value-affirmation or self-superiority priming.

The identical predictions for the value-affirmation and self-superiority priming conditions does not imply that we considered the self-superiority priming procedure fully equivalent to the value-affirmation procedure. As we have explained, however, we had reason to expect that the effects of contemplating one's self-superiority should resemble the effects of contemplating one's values.

Method

Participants and Design

First-year undergraduate students in psychology and social sciences at a university in the Dutch-speaking part of Belgium participated in a study "testing a new evaluation form for courses and professors" (n = 568; 96 men, 470 women, 2 missing values; M_{age} = 18.87 years-old, SD = 2.75, range = 17–56). Although most participants were under 30, the broad age range was due to the presence of five participants who were older than the typical student (these 5 were 32, 36, 39, 40, and 56 years-old). National laws and regulations discourage collecting ethnic background information. However, we asked about participants' native language. Of the 586 participants, 530 (90%) reported that it was Dutch whereas 37 reported another language (1 missing value).

Participants were randomly assigned to a condition of a 3 (initial task: value-affirmation, self-superiority, baseline) by 2 (grade: good, bad) by 2 (professor: man, woman) between-

subjects design. An a priori power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) revealed that we needed 346 participants to observe a medium-size effect with a power of 0.99. We included all students who volunteered to participate in the year when the study was run. The resulting sample size was roughly three times the one of Sinclair and Kunda (2000, Study 1). The study was approved by the Social and Societal Ethical Committee of KU Leuven prior to data collection.

Procedure and Materials

Data collection occurred in groups of 10 to 90. Each participant received a folder with the materials. We randomly distributed the folders over the seats in the classrooms where the data collection took place before we allowed participants to enter and freely take a seat. This procedure guaranteed a random distribution of participants over condition, but precluded the enforcement of equal cell sizes (as varying numbers of seats remained empty in all sessions). For our 3 x 2 x 2 design, cell sizes ranged from 46–49 students.

Oral instructions explained that the folder might include several questionnaires that we had combined for practical reasons, and that participants should fill them out in the order in which they appeared. After having given informed consent, participants worked through the questionnaires at their own pace. Those students in the value-affirmation or self-superiority conditions went through the appropriate procedure before turning to the vignette and the evaluation form; those in the baseline conditions read the vignette right away.

Initial task: Value-affirmation. Select participants rank ordered 15 values (i.e., religiosity, financial security, health, popularity, good family ties, feeling connected with nature, enjoying art and/or music, performing athletically, performing academically and/or professionally, being admired, meeting new people, maintaining lasting friendships, developing

one's talents, becoming a better person, finding the love of one's life) from most important to least important for them personally. They then explained why their personal number one value was important to them, identify a situation where it played a role, and gave up to three examples of when it had guided their actions. The purpose of this task was to prime value-affirmation for students in this condition of the experiment.

Initial task: Self-superiority priming. Select participants rated how much they possessed 26 traits as compared to the average peer of their age, gender, and educational level (bipolar 7-point scales; 1 = *much less*; 7 = *much more*). The traits were chosen such that a majority (but, for believability, not all) would encourage participants to express self-superiority. Specifically, we borrowed the traits from a pretest where 191 student participants (156 women, 35 men, , M_{age} = 18.33 years-old, *SD* = 0.99, range = 17–24) compared the self to others on 12 agentic and 12 communal traits. Within each category, half the traits was desirable; the other half, undesirable. As in most studies on self-superiority beliefs, many more communal than agentic traits reliably yielded a self-other difference. We selected the four positive (helpful, honest, righteous, loving) and four negative (arrogant, untrustworthy, disrespectful, sanctimonious) communal traits that most strongly yielded self-superiority beliefs and added synonyms or closely related terms for each (considerate, sincere, fair, dedicated, pretentious, false, irreverent, hypocritical). We also selected one positive (smart) and three negative agentic (stupid, submissive, ignorant) traits that provoked self-superiority and added a synonym or closely related term for each (perspicacious, decisive, unwise, servile, silly). To address the strong imbalance between positive and negative agentic traits without enhancing the proportion of traits that we did not expect to provoke self-superiority beliefs too much, we added one single pair of positive traits (willful, decisive) that in our pretest had not reliably shown self-superiority.

The resulting self-superiority questionnaire showed satisfactory reliability for both the 16-item communion subscale ($\alpha = .89$) and the 8-item agency subscale ($\alpha = .81$). The purpose of this task was to prime self-superiority for students in this condition.

Vignette. Participants read a description of a hypothetical course taught by "Professor Demuynck" (a name not occurring among the participants' professors). It included details about the course's aim, contents, format, the grading system, the student's obtained grade, and the verbal feedback that the professor had given about that grade. The vignette began by describing the course: "The course 'Basic academic skills' aims at introducing students to scientific language, to the sometimes implicit rules that govern the scientific debate, and to strategies to seek information and to efficiently and properly analyze and synthesize it in collaboration with others."

It then described how Professor Demuynck taught and graded the course. Here the professor's gender was manipulated by repeatedly using the pronouns *he* or *she*. The relevant paragraph read:

> Professor Demuynck teaches the course. He [She] has chosen to do this in the form of seminars. He [She] therefore does not give lectures, but lets the students do a number of assignments. Students choose a topic for these from a list offered by the professor or propose a topic themselves. They do not take a classical exam. Instead, each student must write a paper and give a presentation about the chosen topic. Professor Demuynck has stated that when awarding points, he [she] takes into account the content of the paper, the creativity with which the student has elaborated the topic, and the extent to which deadlines and formal requirements are being met. He [She] also takes into account the quality of the presentation and

the quality of the writing.

Next came a paragraph about how the student had experienced the course:

> You have proposed a subject of your own. Later on, completing the assignment turned out to be more difficult than you had expected because you had to schedule several other demanding courses in the same semester. Fortunately, you have succeeded in completing the written paper on time. You felt that the oral presentation went well, despite the difficult questions being asked by fellow students.

The final paragraph was about the grade and the verbal feedback on it. It included the grade manipulation and a final reminder of the professor's gender. At the participants' university, a grade of 6 (of 20) is a clear fail whereas a grade of 14 yields the mention "distinction." We used these grades in the bad-grade and good-grade condition, respectively. The paragraph read as follows:

> Your final grade was 6 [14]. Professor Demuynck has found your assignments to be clearly not of a sufficiently good level to make you succeed in this course [at the level of distinction]. The professor also said that he [she] got the impression that your work revealed little [great] commitment. Therefore, you will have to resume this course in the next session [have successfully earned your credit for this course].

Evaluation form. The dependent variables were participants' general impressions of the professor and the course. To support the cover story and to check if participants had noticed the information about the grade, the dependent variables were embedded in an extensive survey about the professor, the course, the grade, and the evaluation form itself.

Participants first reported their impression of the professor on an 11-point response scale from 0 (*extremely bad*) to 10 (*outstanding*). They then indicated their agreement with statements that described the professor on 12 traits, half positive and half negative, and half warmth-related (positive: accessible, understanding, friendly; negative: strict, condescending, pompous) and half competence-related (positive: expert, purposive, energetic; negative: unorganized, lacking foresight, lazy). They also responded to the statements: "The professor is suitable for the topic being taught" and "The professor is suitable for the course format being chosen." Participants gave all these specific judgments on 7-point response scales from 1 (*totally disagree*) to 7 (*totally agree*). The internal consistency of the 14-item professor impression scale was good (Cronbach's alpha after reversed coding of the negative items = .89). Although the scale merely served to support our cover story, scores were averaged across items so that higher scores indicate a more favorable impression of the professor.

Participants then reported their general impression of the course, again on an 11-point response scale from 0 (*extremely bad*) to 10 (*outstanding*). After that, they indicated their agreement with statements that described six course characteristics (important, inspiring, meaningful, difficult, unnecessary, boring), and the statement: "The chosen course format is suitable for the topic being taught." Participants expressed their agreement with each of the eight statements on a 7-point response scale from 1 (*totally disagree*) to 7 (*totally agree*). The internal consistency of the course impression scale that was based on these specific course judgments was good (Cronbach's alpha Cronbach's alpha after reversed coding of the negative items = .69). Although this scale, too, served to support our cover story, responses were averaged across items so that higher scores signified a more positive impression of the course.

The form also included questions about the grade and the manner in which it had been awarded. Participants expressed their satisfaction with the described grade on a scale from 0 (*extremely dissatisfied*) to 10 (*extremely satisfied*). They then gave a general judgment of how grades were awarded, judged the manner in which grades were awarded on four characteristics (just, expert, careful, applying clear criteria), and responded to the statement: "The chosen exam format is suitable for the topic being taught." Participants also rated emotions that the grade would make them experience. Because of an error in the instruction for these ratings, we could not use these data. The internal consistency of the specific judgments of the six-item grade system scale was good (Cronbach's $\alpha = 0.84$). Scores across items were averaged so that higher scores indicated stronger satisfaction with the awarded grade.

Finally, participants judged the evaluation form by responding to statements about their willingness to fill out the form in the future and about the extent to which the questions reflected the most relevant aspects of teachers, course contents, course formats, and grading systems. Participants expressed their agreement with those positively worded statements on 7-point response scales from 1 (*totally disagree*) to 7 (*totally agree*); $\alpha = .71$). The purpose of these items was to support the stated rationale for the data collection.

The questionnaire also included open filler questions interspersed between the various scales. Four questions asked about the characteristics that a professor should have to be able to teach the topic and use the format of the course being described as well as ideal course and exam formats. Four questions asked which characteristics or aspects of professors, course contents, course formats, and exam formats were missing or unduly present in the form. We needed these open questions to uphold the cover story and enhance the form's face validity because SETs at participants' university typically allow students to write down comments. However, using a word

count, we checked how extensively participants had answered the open questions as an indication of how attentively and seriously they had filled out the questionnaire.

Results

In a preliminary analysis on the impressions of the course and the professor, participants' gender was not involved in any interaction, including interactions with professor gender. Because of that, and of the highly uneven number of men and women, we limited the number of effects being tested (thus avoiding the multiple comparisons problem) by dropping participants' gender from the analyses. One participant in the value-affirmation condition did not order all values, and four participants in the self-superiority condition did not describe themselves as superior to others on communion nor agency. Because these participants may not have self-affirmed, we analyzed the data one time including and a second time excluding them. The analyses yielded nearly identical results, with no changes in significances. We report the results for the full dataset (those for the analysis without the five participants are presented in the online supplement). Differing degrees of freedom are due to missing values. All confidence intervals are confidence intervals of the difference.

Manipulation Checks

We checked if the grade manipulation was successful by analyzing participants' satisfaction scores, participants' general impressions of the grading system, and participants' impressions as derived from the grading system scale. Besides grade, we included the independent variables of initial task and professor to allow us to assess if the grade manipulation was equally successful across the conditions of the other independent variables. Each 3 x 2 x 2 ANOVA thus included initial task (value-affirmation, self-superiority, baseline), grade (good, bad), and professor (man, woman) as between-subjects variables.

All analyses showed a main effect of grade. Participants were more satisfied with a good grade (M = 8.23, SD = 1.52) than with a bad grade (M = 1.97, SD = 1.54), $F(1,535) = 2293.79$, $p < .001$, $\eta_p^2 = .81$. In addition, general impressions of the grading system were better in the good-grade condition (M = 6.81, SD = 1.64) than in the bad-grade condition (M = 4.32, SD = 1.55), $F(1,552) = 340.52$, $p < .001$, $\eta_p^2 = .38$, and specific judgments of the grading system were also better in the good-grade condition (M = 4.89, SD = 0.86) than in the bad-grade condition (M = 4.08; SD = 0.92), $F(1,554) = 116.23$, $p < .001$, $\eta_p^2 = .17$.

Of all participants, 505 (89%) wrote at least one comment. Of these, 458 (81%) wrote about the characteristics of a good professor, 332 (58%) about the format of the course or the exam, and 338 (60%) about elements of the evaluation form. An ANOVA on the number of words written revealed no effects of our manipulations; all Fs < 2.04, ps > .153. It seems, then, that participants were generally motivated to fill out the questionnaire as well as they could.

Tests of Hypotheses

As in Sinclair and Kunda (2000, Study 1), participants' general impressions of the professor and the course were correlated (r = .52, $p < .001$). We averaged them and report analyses on the average scores. Separate analyses revealed identical patterns.

Hypothesis 1. We predicted a two-way interaction of grade and professor gender in that students in the bad-grade condition, but not those in the good-grade condition, would give lower SETs to the female professor than to the male professor. To test this hypothesis, we subjected the impression scores in the baseline condition to a 2 x 2 ANOVA with grade (good, bad) and professor (man, woman) as between-subjects variables and combined course/professor impressions as the dependent variable.

We found a main effect of grade, $F(1,181) = 96.74$, $p < .001$, $\eta_p^2 = .35$. Participants reported better impressions after a good grade ($M = 6.87$, $SD = 0.95$) than after a bad grade ($M = 5.32$, $SD = 1.27$). The main effect of professor was not significant, $F(1,181) = 2.70$, $p = .102$, $\eta_p^2 = .02$. Importantly, and in support of Hypothesis 1, grade interacted with professor, $F(1,181) = 9.87$, $p = .002$, $\eta_p^2 = .05$. The cell means are presented in Figure 1. Participants gave a higher evaluation in the good-grade condition than in the bad-grade condition, but the difference was larger if the professor was a woman, $t(88) = 9.17$, $p < .001$, $d = 1.91$, 95% CI [-2.52, -1.62], than if the professor was a man, $t(93) = 4.74$, $p < .001$, 95% CI [-1.52, -0.62], $d = 0.98$. Consistent with Hypothesis 1, participants with a good grade did not judge the professors differently, $t(94) = 1.24$, $p = .218$, $d = 0.25$, 95% CI [-0.62, 0.14], whereas participants in the bad-grade condition gave a higher SET if the professor was a man rather than a woman, $t(87) = 2.96$, $p = .004$, $d = 0.63$, 95% CI [0.25, 1.28]. We thus replicated Sinclair and Kunda (2000, Study 1) by showing a gender bias in the bad-grade condition but not in the good-grade condition. The stage was set to examine the effect of value-affirmation and self-superiority priming on the gender bias.

Hypothesis 2. We predicted that value-affirmation and self-superiority priming would eradicate the difference that occurred in the bad-grade condition between the evaluations of male and female professors. We thus predicted a three-way interaction of initial task, grade, and professor gender. To test the prediction, we conducted a 3 x 2 x 2 ANOVA with initial task (value-affirmation, self-superiority, baseline), grade (good, bad), and professor (man, woman) as between-subjects variables and combined course/professor impressions as the dependent variable.

The main effect of grade remained significant, $F(1,533) = 292.48$, $p < .001$, $\eta_p^2 = .35$. Participants gave higher evaluations in the good-grade condition ($M = 6.87$, $SD = 1.05$) than in the bad-grade condition ($M = 5.09$, $SD = 1.40$). A main effect of initial task also occurred, $F(2,533) =$

3.62, $p = .027$, $\eta_p^2 = .01$, but the pairwise contrasts were not significant (Tukey $ps \geq .06$). The main effect of professor was not significant, $F(1,533) = 2.41$, $p = .121$, $\eta_p^2 < .01$, nor was the two-way interaction of grade and professor, $F(1,533) = 0.05$, $p = .824$, $\eta_p^2 < .01$.

Importantly, and in support of Hypothesis 2, the predicted three-way interaction of initial task, grade, and professor was significant, $F(2,533) = 5.63$, $p = .004$, $\eta_p^2 = .02$. The cell means and standard deviations are presented in Figure 2. Our hypothesis predicted an interaction of professor and initial task in the bad-grade condition, but not in the good-grade condition. We therefore broke the three-way interaction down per grade.

In the good-grade condition, consistent with the pattern predicted by Hypothesis 2, the professor by initial task interaction was not significant, $F(1,271) = 2.35$, $p = .098$, $\eta_p^2 = .02$, nor were the main effects of professor, $F(1,271) = 1.23$, $p = .268$, $\eta_p^2 < .01$, and initial task, $F(1,271) = 2.26$, $p = .107$, $\eta_p^2 = .02$. Participants did not differentiate between the male and the female professor at baseline, in the value affirmation condition, $t(89) = 1.55$, $p = .125$, $d = 0.33$, 95% CI [-0.09, 0.73], or in the self-superiority condition, $t(88) = 1.34$, $p = .178$, $d = 0.29$, 95% CI [-0.16, 0.83] (see Figure 2a). In sum, in the good-grade condition, male and female professors were rated similarly across all three initial task conditions.

In the bad-grade condition, consistent with the pattern predicted by Hypothesis 2, the two-way interaction of professor and initial task was significant, $F(1,262) = 3.45$, $p = .033$, $\eta_p^2 = .03$. Participants with a bad grade showed a gender bias at baseline, but not after value affirmation, $t(84) = 1.17$, $p = .245$, $d = 0.25$, 95% CI [-0.86, 0.22], or self-superiority priming, $t(91) = 0.33$, $p = .743$, $d = 0.07$, 95% CI [-0.55, 0.77] (see Figure 2b). In sum, in the bad-grade condition, the lower rating of the female compared to the male professor in the baseline condition was not evident in either priming condition.

We also explored the three-way interaction by examining how the initial task affected participants' impressions per grade and within professor gender. One-way ANOVAs showed an effect of initial task in the bad-grade/male-professor condition only, $F(2,133) = 4.43$, $p = .014$, $\eta_p^2 = .06$. Participants gave higher SETs at baseline than after value-affirmation (Tukey $p = .036$, $d = 0.49$) or self-superiority priming (Tukey $p = .026$, $d = 0.51$), with the latter conditions not differing from each other (Tukey $p = 0.994$) (see Figure 2b). In other words, in the bad-grade condition, male professors were rated more positively in the baseline condition than in either of the priming conditions. The initial task did not affect impression scores in any other condition: the bad-grade/female-professor condition, $F(2,129) = 1.40$, $p = 0.251$, $\eta_p^2 = .02$, the good-grade/male-professor condition, $F(2,136) = 2.45$, $p = .090$, $\eta_p^2 = .03$, and the good-grade/female professor: $F(2,135) = 2.14$, $p = .122$, $\eta_p^2 = .03$.

In summary, value-affirmation and self-superiority priming eradicated gender bias in the bad-grade condition. The more equal impressions were due to SETs for the male professor becoming lower rather than to SETs for the female professor becoming higher.

Exploratory Analyses

Cover story. We did not predict any effects on the scales bolstering the cover story (specific judgments and judgments of the evaluation form), but we analysed them exploratorily. Mean ratings of the course and the professor were correlated ($r = .42$, $p < .001$). We averaged them and subjected the average scores to an ANOVA with the same design as the ANOVA on our dependent variables. We found a main effect of grade, $F(1,556) = 281.27$, $p < .001$, $\eta_p^2 = .34$. Participants gave higher ratings after a good grade ($M = 4.69$, $SD = 0.48$) than after a bad grade ($M = 3.98$, $SD = 0.53$). No other effects were significant (Fs < 2.48 ps > .085). We also analyzed average ratings on the items about the evaluation form. The ANOVA yielded a main effect of grade, $F(1,552) = 4.81$, $p = .029$,

η_p^2 = .01. Participants found the evaluation form better after a good grade (M = 5.11, SD = 0.82) than after a bad grade (M = 4.96, SD = 0.74). No other effects were significant (Fs < 2.56, ps > .078).

Correlations. At many universities, students fill out SET forms voluntarily. If their willingness to do so is associated with their appreciation of male and female professors, that may in itself be a hidden source of bias. We therefore explored how participants' impressions of the course and the professor were related to their attitudes toward the evaluation form. We did so separately for the baseline condition (where gender bias occurred) and the self-affirmation conditions combined (where no gender bias occurred), as well as separately for the male-professor and female-professor conditions.

In the baseline conditions, the positivity of participants' attitude toward the evaluation form was correlated with the positivity of their impression of the course and the professor. The correlation was significant if the professor was a woman (r = .25, p = .019) and not if the professor was a man (r = .12, p = .234), but the difference between the two correlations was not significant (z = 0.90, p = .368). In the self-affirmation conditions, the correlation between participants' attitude toward the evaluation form and their impression of the course and the professor was stronger than in the baseline condition if the professor was a man (r = .36, p < .001; z = 2.17, p = .048) and lower than in the baseline condition if the professor was a woman (r = .03, p = .710; z = 1.72, p = .085). As a result, the correlation between participants' attitude toward the evaluation form and their impression of the course and the professor was higher if the professor was a man than if she was a woman (z = 3.23, p = .001). Eradicating the gender bias in the content of the SET thus went hand in hand with introducing a gender bias in participants' attitudes toward the SET task itself.

Discussion

We tested the effect of self-affirmation on gender bias in Belgian students' evaluations of teaching. Participants in the bad-grade condition and who had not gone through a self-affirmation procedure favored the male over the female professor. This result provides converging evidence for the robustness of the finding in Sinclair and Kunda (2000, Study 1), who demonstrated this grade-contingent gender bias using another methodology and on students from another country.

Contemplating one's values or one's personal superiority eradicated the gender bias in the evaluation of the course and the professor (but, as we will explain, introduced a kind of gender bias in attitudes toward the SET itself). This finding extends earlier research that has shown that value-affirmation reduced prejudice based on ethnicity or sexual orientation (Fein & Spencer, 1997; Lehmiller et al., 2010) in the absence of self-threat (Lehmiller et al., 2010) or under self-threat coming from a third person (Fein & Spencer, 1997). The finding that self-affirmation reduced the gender bias by rendering the evaluation of the male professor lower suggests that the gender bias at baseline involved an overly high evaluation of the male professor rather than, as suggested by Sinclair and Kunda (2000), a derogation of the female professor.

Limitations and Future Research Directions

Most of our participants were women. That limitation was unavoidable given students' study programs but may cast doubt on the generalizability of our findings. However, earlier research suggests that gender bias is larger among male than among female students (Boring, 2017; Mengel et al., 2019). If anything, therefore, our study underestimates rather than overestimates the gender bias that would occur in mixed-gender groups of students who have received a bad grade and have not self-affirmed.

As we explained in the introduction, our choice of a vignette approach was mainly inspired by legal and ethical considerations. Besides avoiding these problems, our approach had the advantage that it gave us full experimental control over the independent variables and potential confounds. Still, one may wonder if the gains in terms of legal and ethical acceptability and internal validity warrant the loss of ecological validity that is potentially associated with a vignette approach. One easy answer would be that addressing a question through an approach with limitations is preferable to not being able to examine that question at all. Yet, we also had more positive reasons to trust our results.

One reason was that in earlier studies, the occurrence of a gender bias in SETs did not depend on the research approach being used (compare, for example, Kierstead et al., 1988; MacNell et al., 2015; Wagner et al., 2016). At least one study also used evaluations of totally hypothetical professors (Kierstead et al., 1988). The consistency of findings across methods provides convergent evidence of the validity of vignette studies. Another reason to trust our findings only became evident after the data analysis. The baseline conditions of our design conceptually replicated several real-life studies on grade effects (Ewing, 2012; Gorry, 2017; Weinberg et al., 2009) and a real-life study on a combined grade and gender bias in SETs (Sinclair & Kunda, 2000, Study 1). We indeed found that students who imagined a good grade gave higher evaluations than students who imagined a bad grade and that only those who imagined a bad grade showed a gender bias.

One issue that all researchers conducting vignette studies need to address is striking a balance between giving participants sufficient details and presenting them with digestible amounts of information. One extreme would be to try to simulate the richness of information that students normally have available when providing SETs after having taken semester-long real-life

courses. This approach would suggest presenting participants with detailed syllabi, lecture or assignment outlines, excerpts of reading materials, and perhaps even audio or video recordings of classes. Because that would arguably render the experimental task daunting for most participants, it would entail considerable variation in the attention paid to experimental materials, as well as substantial self-selection. An even more serious problem is that priming participants with detailed information might reduce, rather than enhance, ecological validity. It is probably safe to assume that hardly any student reviews all the details of the written and oral communications about a course while filling out real-life SETs. We therefore opted for vignettes that included brief descriptions of the course, the professor, and the grading system.

Given that we studied simulated SETs using a vignette approach, however, one might wonder how self-affirmation would affect real-life SETs that students provide after having followed a course during a full semester. However, it is important to emphasize that our research did replicate some key findings from the SET literature. We therefore argue that our research is relevant to actual SETs. Still, it would be interesting to replicate our research using different methodologies. Such follow-up research may also address another limitation of our research, that is, that the self-affirmation procedures occurred just before participants learned about and evaluated the course. This timing raises a question about whether the self-affirmation procedures altered the extent to which the participants experienced the self-threatening information as threatening, whether it modified the manner in which they coped with an identical threat, or both.

We found that self-affirmation affected the association between the positivity of the evaluation and attitudes toward the evaluation itself differently for male and female professors. However, the finding should be treated with caution because it emerged in a purely exploratory analysis. Moreover, it is unclear if students' attitudes toward the evaluation predicted their

willingness to submit SETs. Future research testing the effect of self-affirmation on attitudes toward SETs and on the relationship between these attitudes and students' willingness to provide sets will contribute to a fuller view of the effect of self-affirmation on SETs.

Practice Implications

Despite its limitations, our research has implications for schools and universities that wish to de-bias SETs. At first sight, the finding that self-affirmation eradicated the gender bias may seem a good tiding. All it seems to take to eradicate bias is to make students contemplate their values or their personal superiority before giving SETs. A brief self-superiority priming questionnaire preceding a SET form may be particularly well-suited for this goal because first having to judge one's own personal qualities before evaluating an activity that serves to contribute to the development of these qualities (that is, a course taken) arguably has great face validity in the eyes of students.

Yet, the finding that self-affirmation affects the association between students' evaluation of a course-professor combination and their attitude toward the SET itself calls for caution. If attitudes toward SETs affect students' willingness to fill them out in the first place, it may imply that students become more willing to provide SETs for male professors whose teaching they appreciate, whereas such an effect does not take place for female professors. Although eradicating gender bias in the contents of SETs, self-affirmation may enhance gender bias in the likelihood that SETs are being submitted. If this interpretation of our findings is valid (something that follow-up research will have to examine), then self-affirmation procedures may above all be advisable in contexts where all students routinely fill out SETs (rather than only those who volunteer to do so).

The finding that the gender bias at baseline occurred only among students who had obtained a bad grade suggests a manner to unequivocally detect it in actual SETs. We recommend that schools and universities not just check for an overall gender bias, but also for a gender bias as a function of students' results. The finding that students with poor grades, but not those with good grades, differentiate between male and female professors would provide unequivocal evidence for a gender bias rather than some "real" gender difference in teaching quality within a given school or program. As such, it would arguably provide a basis for statistical correction that would be more acceptable in the eyes of the instructor team than generally enhancing scores for women or reducing scores for men. Alternatively, schools and programs might consider eliminating the SET scores from students with bad grades as an instrument for academic decisions concerning professors from that program or school.

Of course, our recommendation only works if students' grades can be matched with the SETs they provide (which is impossible when SETs are anonymous) or when students honestly report their grades on the SET form. When neither condition is fulfilled, the best approach may be to collect SETs before students learn about their grades and perhaps even before exams take place. The downside of this approach is that it does not allow students to incorporate their experience with the exam. Ideally, therefore, SETs should be invited in two parts: one part before and one part after students have taken the exam and learned about their grades.

More generally, our findings have implications for schools and universities that include SETs in hiring and promotion decisions. As we explained in the introduction, SETs are mostly unrelated to more objective measures of teaching effectiveness. If they are related to such measures, it is often negatively rather than positively (Boring et al., 2016; Braga et al., 2014; Carrell & West, 2010). The accumulation of evidence that SETs are systematically biased,

together with their non-existent or negative relationship with teaching effectiveness, may be an argument to strictly limit their weight in decisions affecting instructors' careers.

Theoretical Implications

The finding that self-affirmation reduced the gender bias by rendering the evaluation of the male professor lower nuances the prevailing view that self-affirmation has virtually uniquely beneficial interpersonal effects, such as greater acceptance of criticism (Van Tongeren et al., 2014) and more openness to other people's views (Binning, Sherman, & Cohen, 2010). Interestingly, a similar finding has occurred in at least one self-affirmation study (Cohen et al., 2000, Study 3). In that research, self-affirmation eradicated partisan bias by rendering participants' judgment of an individual who supported their attitude more harsh rather than by rendering their judgment of an individual who opposed it more lenient.

We do not suggest that self-affirmation-induced reductions of prejudice are always due to judgments of advantaged groups turning less favorable. What we do infer from our and Cohen et al.'s (2000) findings is that the reduction of bias and prejudice toward a disadvantaged group *may* completely or partially take the form of judgments of a advantaged group getting less positively biased. If proven valid, this inference may call for a reconsideration of the nature of self-affirmation effects on interpersonal judgments and for more attention to what some authors have called "the dark side of self-affirmation" (Munro & Stansbury, 2009, p. 1143).

Some recent studies have indeed revealed circumstances where self-affirmation negatively affects cognitive biases (Munro & Stansbury, 2009), stress (Jessop, Ayers, Burn, & Ryda, 2018) and openness to health recommendations (Ferrer, Klein, & Graff, 2017). Our research thus contributes to a small but growing body of evidence that self-affirmation is not all positive. To fully understand self-affirmation effects on prejudice, therefore, future studies would

do well including measures of how participants view all groups involved rather than merely examining intergroup differences or participants' views of the disadvantaged group.

Conclusion

Students evaluate the teaching of professors lower after having received a bad grade from them than after a good grade. After a bad grade, they show a gender bias in that they evaluate the course of a male professor more leniently than the course of a female professor. Value-affirmation and self-superiority priming reduce this gender bias by lowering the evaluation of the male professor rather than by enhancing the evaluation of the female professor. However, both self-affirmation procedures inflate the positive correlation between students' evaluation of teaching and their attitude toward the evaluation procedure itself for a male professor while eradicating it for a female professor. The gender bias may thus return in the form of a greater willingness to fill out a student evaluation of teaching for highly appreciated courses of male (vs. female) professors. Although the present findings suggest ideas for reducing the gender bias (and perhaps other biases as well) in student evaluations of teaching, they also contribute to the growing body of evidence that using those evaluations in hiring and promotion decisions of faculty is highly problematic.

References

Abel, M. H., & Meltzer, A. L. (2007). Student ratings of a male and female professors' lecture on sex discrimination in the workforce. *Sex Roles, 57,* 173–180. https://doi.org/10.1007/s11199-007-9245-x

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49,* 1621–1630. https://doi.org/10.1037/0022-3514.49.6.1621

Allen, T. J., & Sherman, J. W. (2011). Ego threat and intergroup bias: A test of motivated-activation versus self-regulatory accounts. *Psychological Science, 22,* 331–333. https://doi.org/10.1177/0956797611399291

Arbuckle, J., & Williams, B. D. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles, 49,* 507–516. https://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,* 193–210. https://doi.org/10.1080/03634529909379169

Badea, C., & Sherman, D. K. (2019). Self-Affirmation and prejudice reduction: When and why ? *Current Directions in Psychological Science, 28,* 40–46. https://doi.org/10.1177/0963721418807705

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

Bavishi, A., Madera, J. M., & Hebl, M. R. (2010). The effect of professor ethnicity and gender on student evaluations: Judged before met. *Journal of Diversity in Higher Education, 3,* 245–256. https://doi.org/10.1037/a0020763

Binning, K. R., Sherman, D. K., & Cohen, G. L. (2010). Seeing the other side : Reducing political partisanship via self-affirmation in the 2008 presidential election. *Analysis of Social Issues and Public Policy, 10,* 276–292. https://doi.org/10.1111/j.1530-2415.2010.01210.x

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145,* 27–41. https://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,* 1–11. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41,* 71–88. https://doi.org/10.1016/j.econedurev.2014.04.002

Brown, J. D. (1986). Self-enhancement biases in social judgments. *Social Cognition, 4,* 353–376. https://doi.org/10.1521/soco.1986.4.4.353

Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin, 38,* 209–219. https://doi.org/10.1177/0146167211432763

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118,* 409–432. https://doi.org/10.1086/653808

Carrell, S. E., & West, J. E. (2016). Does professor quality matter ? Evidence from random assignment of students to professors. *Journal of Political Economy*, *118*, 409–432. https://doi.org/10.1086/653808

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

Chien, Y.-W., Wegener, D. T., Petty, R. E., & Hsiao, C.-C. (2014). The Flexible Correction Model: Bias correction guided by naïve theories of bias. *Social and Personality Psychology Compass*, *8*, 275–286. https://doi.org/10.1111/spc3.12105

Clayson, D. E. (2019). Student perception of instructors: The effect of age, gender and political leaning and political leaning. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2019.1679715

Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality & Social Psychology Bulletin*, *26*, 1151–1164. https://doi.org/10.1177/01461672002611011

Cohen, G. L., & Sherman, D. K. (2014). The psychology of change: Self-affirmation and social psychological intervention. *Annual Review of Psychology*, *65*, 333–371. https://doi.org/10.1146/annurev-psych-010213-115137

Collange, J., Benbouzyane, L., & Sanitioso, R. (2006). Self-image maintenance and discriminatory behavior. *Revue Internationale de Psychologie Sociale*, *19*(3–4), 153–171. Retrieved from https://www.cairn.info/revue-internationale-de-psychologie-sociale-2006-3-page-153.htm

Collange, J., Fiske, S. T., & Sanitioso, R. (2009). Maintaining a positive self-image by stereotyping others: Self-threat and the stereotype content model. *Social Cognition, 27,* 138–149. https://doi.org/10.1521/soco.2009.27.1.138

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin, 129,* 414–446. https://doi.org/10.1037/0033-2909.129.3.414

Critcher, C. R., & Dunning, D. (2015). Self-affirmations provide a broader perspective on self-threat. *Personality and Social Psychology Bulletin, 41,* 3–18. https://doi.org/10.1177/0146167214554956

Echterhoff, G., Groll, S., & Hirst, W. (2007). Tainted truth: Overcorrection for misinformation influence on eyewitness memory. *Social Cognition, 25,* 367–409. https://doi.org/10.1521/soco.2007.25.3.367

Epton, T., Harris, P. R., Kane, R., Koningsbruggen, G. M. Van, & Sheeran, P. (2015). The impact of self-affirmation on health-behavior change: A meta-analysis. *Health Psychology, 34,* 187–196. https://doi.org/10.1037/hea0000116

Ewing, A. M. (2012). Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review, 31,* 141–154. https://doi.org/10.1016/j.econedurev.2011.10.002

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PloS one, 14,* e0209749. https://doi.org/10.1371/journal.pone.0209749

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research*

Methods, *39*, 175–191. https://doi.org/10.3758/BF03193146

Fein, S., & Spencer, S. J. (1997). Prejudice as self-image maintenance: Affirming the self through derogating others. *Journal of Personality and Social Psychology*, *73*, 31–44. https://doi.org/10.1037/0022-3514.73.1.31

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*, 317–375. Retrieved from http://www.jstor.org/stable/40196030

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*, 151–211. Retrieved from http://www.jstor.org/stable/40196104

Ferrer, R. A., Klein, W. M. P., & Graff, K. A. (2017). Self-affirmation increases defensiveness toward health risk information among those experiencing negative emotions: Results from two national samples. *Health Psychology*, *36*, 380–391. https://doi.org/10.1037/hea0000460

Fiedler, K., & Semin, G. R. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology*, *54*, 558–568. https://doi.org/10.1037/0022-3514.54.4.558

Fisher, A. N., Stinson, D. A., & Kalajdzic, A. (2019). Unpacking backlash: Individual and contextual moderators of bias against female professors. *Basic and Applied Social Psychology*, *41*, 305–325. https://doi.org/10.1080/01973533.2019.1652178

Gorry, D. (2017). The impact of grade ceilings on student grades and course evaluations: Evidence from a policy change. *Economics of Education Review*, *56*, 133–140. https://doi.org/10.1016/j.econedurev.2016.12.006

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52,* 1209–1217. https://doi.org/10.1037//0003-066X.52.11.1209

Harvey, R. D., & Oswald, D. L. (2000). Collective guilt and shame as motivation for white support of black programs. *Journal of Applied Social Psychology, 30,* 1790–1811. https://doi.org/10.1111/j.1559-1816.2000.tb02468.x

Hoorens, V. (1990). Nuttin's affective selfparticles hypothesis and the name letter effect: A review. *Psychologica Belgica*, 30, 23-48.

Hoorens, V. (2014). What's really in a Name-Letter Effect? Name-letter preferences as indirect measures of self-esteem. *European Review of Social Psychology, 25,* 228–262. https://doi.org/10.1080/10463283.2014.980085

Hoorens, V., & Harris, P. R. (1998). Distortions in reports of health behaviors: The time span effect and illusory superiority. *Psychology and Health, 13,* 451–466. https://doi.org/10.1080/08870449808407303

Hoorens, V., Smits, T., & Shepperd, J. A. (2008). Comparative optimism in the spontaneous generation of future life-events. *British Journal of Social Psychology, 47,* 441–451. https://doi.org/10.1348/014466607X236023 |

Jessop, D. C., Ayers, S., Burn, F., & Ryda, C. (2018). Can self-affirmation exacerbate adverse reactions to stress under certain conditions? *Psychology and Health, 33,* 827–845. https://doi.org/10.1080/08870446.2017.1421187

Joye, S., & Wilson, J. H. (2015). Professor age and gender affect student perceptions and grades. *Journal of the Scholarship of Teaching and Learning, 15,* 126-138. https://doi.org/10.14434/josotl.v15i4.13466

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,* 342–344. https://doi.org/10.1037/0022-0663.80.3.342

Krautmann, A. C., & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review, 18,* 59–63. https://doi.org/10.1016/S0272-7757(98)00004-1

Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice. *Psychological Science, 22,* 1472–1477. https://doi.org/10.1177/0956797611427918

Lehmiller, J. J., Law, A. T., & Tormala, T. T. (2010). The effect of self-affirmation on sexual prejudice. *Journal of Experimental Social Psychology, 46,* 276–285. https://doi.org/10.1016/j.jesp.2009.11.009

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40,* 291–303. https://doi.org/10.1007/s10755-014-9313-4

Mendes, W. B., & Koslov, K. (2013). Brittle smiles: Positive biases toward stigmatized and outgroup targets. *Journal of Experimental Psychology: General, 142,* 923–933. https://doi.org/10.1037/a0029663

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association, 17,* 535–566. https://doi.org/10.1093/jeea/jvx057

Miron, A. M., & Brehm, J. W. (2006). Reactance theory - 40 years later. *Zeitschrift Fur Sozialpsychologie, 37,* 9–18. https://doi.org/10.1024/0044-3514.37.1.9

Mühlberger, C., & Jonas, E. (2019). Reactance Theory. In K. Sassenberg & M. L. W. Vliek (Eds.), *Social psychology in action. Evidence-based interventions from theory to practice.*

(pp. 79–94). Cham: Springer.

Munro, G. D., & Stansbury, J. A. (2009). The dark side of self-affirmation: Confirmation bias and illusory correlation in response to threatening information. *Personality and Social Psychology Bulletin*, *35*, 1143–1153. https://doi.org/10.1177/0146167209337163

Nadler, J. T., Berry, S. A., & Stockdale, M. S. (2013). Familiarity and sex based stereotypes on instant impressions of male and female faculty. *Social Psychology of Education*, *16*, 517-539. https://doi.org/10.1007/s11218-013-9217-7

Napper, L., Harris, P. R., & Epton, T. (2009). Developing and testing a self-affirmation manipulation. *Self and Identity*, *8*, 45–62. https://doi.org/10.1080/15298860802079786

Nuttin, J. M. (1985). Narcissism beyond Gestalt and awareness. *European Journal of Social Psychology*, *15*, 353–361. https://doi.org/10.1002/ejsp.2420150309

Nuttin Jr, J. M. (1987). Affective consequences of mere ownership: The name letter effect in twelve European languages. *European Journal of Social Psychology*, *17*, 381-402. https://doi.org/10.1002/ejsp.2420170402

Peterson, D. A. M., Biederman, L. A., Andersen, D., Ditonto, T. M., & Roe, K. (2019). Mitigating gender bias in student evaluations of teaching. *PLoS ONE*, *14*, 1–10. https://doi.org/10.1371/journal.pone.0216241

Petty, Richard E., Wegener, D.T., White, P. H. (1998). Flexible correction processes in social judgment: Implications for persuasion. *Social Cognition*, *16*, 93–113.

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile?. *Quality Assurance in Education*, *15*, 178-191. https://doi.org/10.1108/09684880710748938

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors. Com. *Journal of Diversity in higher Education*, *3*, 137-152.

https://doi.org/10.1037/a0019865

Schumann, K. (2014). An affirmed self and a better apology: The effect of self-affirmation on transgressors' responses to victims. *Journal of Experimental Social Psychology*, *54*, 89–96. https://doi.org/10.1016/j.jesp.2014.04.013

Seta, J. J., Seta, C. E., & McCormick, M. (2020). An overcorrection framing effect. *Journal of Behavioral Decision Making*, *33*, 27–38. https://doi.org/10.1002/bdm.2143

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 38, pp. 183–242). San Diego, CA: Academic Press. https://doi.org/10.1016/S0065-2601(06)38004-5

Sherman, D. K., & Hartson, K. A. (2011). Reconciling self-protection with self-improvement: Self-affirmation theory. In M. D. Alicke & C. Sedikides (Eds.), *Handbook of self-enhancement and self-protection* (pp. 128–151). New York: Guilford Press.

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

Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin*, *26*, 1329–1342. https://doi.org/10.1177/0146167200263002

Sommers, S. R., & Kassin, S. M. (2001). On the many impacts of inadmissible testimony: Selective compliance, need for cognition, and the overcorrection bias. *Personality and Social Psychology Bulletin*, *27*, 1368–1377. https://doi.org/10.1177/01461672012710012

Stone, J., Whitehead, J., Schmader, T., & Focella, E. (2011). Thanks for asking: Self-affirming questions reduce backlash when stigmatized targets confront prejudice. *Journal of*

Experimental Social Psychology, *47*, 589–598. https://doi.org/10.1016/j.jesp.2010.12.016

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science*, *11*, 800-816. https://doi.org/10.1177/1745691616650284

Van Tongeren, D. R., Green, J. D., Hulsey, T. L., Legare, C. H., Bromley, D. G., & Houtman, A. M. (2014). A meaning-based approach to humility: Relationship affirmation reduces worldview defense. *Journal of Psychology and Theology*, *42*, 62–69. https://doi.org/10.1177/009164711404200107

Wagner, N., Rieger, M., & Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review*, *54*, 79–94. https://doi.org/10.1016/j.econedurev.2016.06.004

Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *The Journal of Economic Education*, *40*, 227–261. https://doi.org/10.3200/JECE.40.3.227-261

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*, 806–820. https://doi.org/10.1037//0022-3514.39.5.806

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. *Assessment and Evaluation in Higher Education*, *37*, 683–699. https://doi.org/10.1080/02602938.2011.563279

Youmans, R. J., & Jee, B. D. (2007). Fudging the numbers: Distributing chocolate influences student evaluations of an undergraduate course. *Teaching of Psychology*, *34*, 245–247. https://doi.org/10.1080/00986280701700318

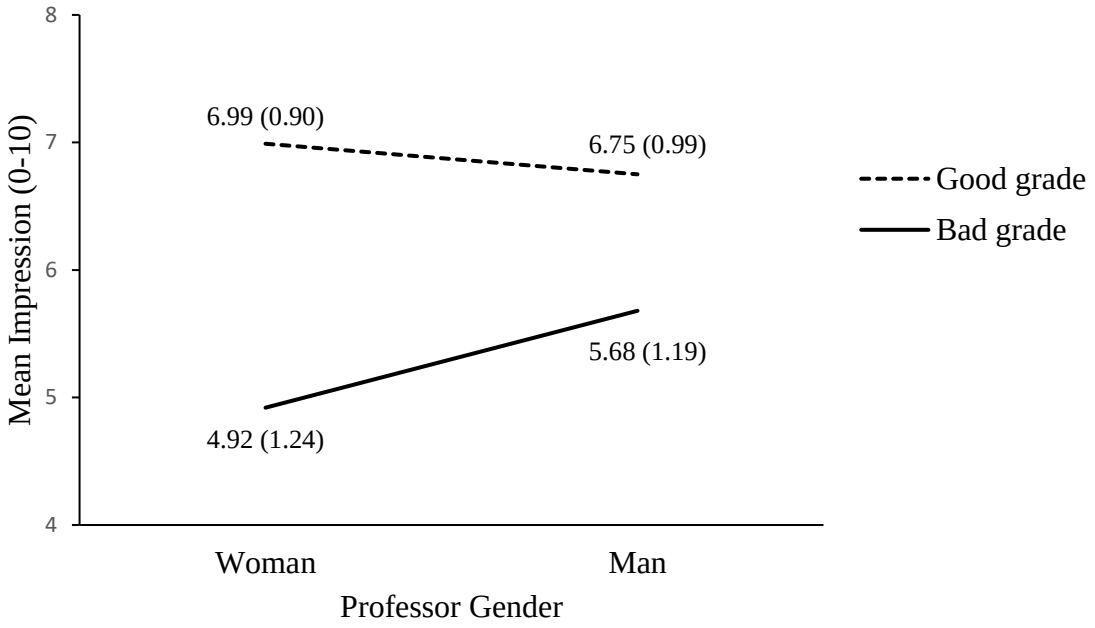

Figure 1. The interactive effect of grade and professor gender on impressions in the baseline condition (*SD*s in parentheses).

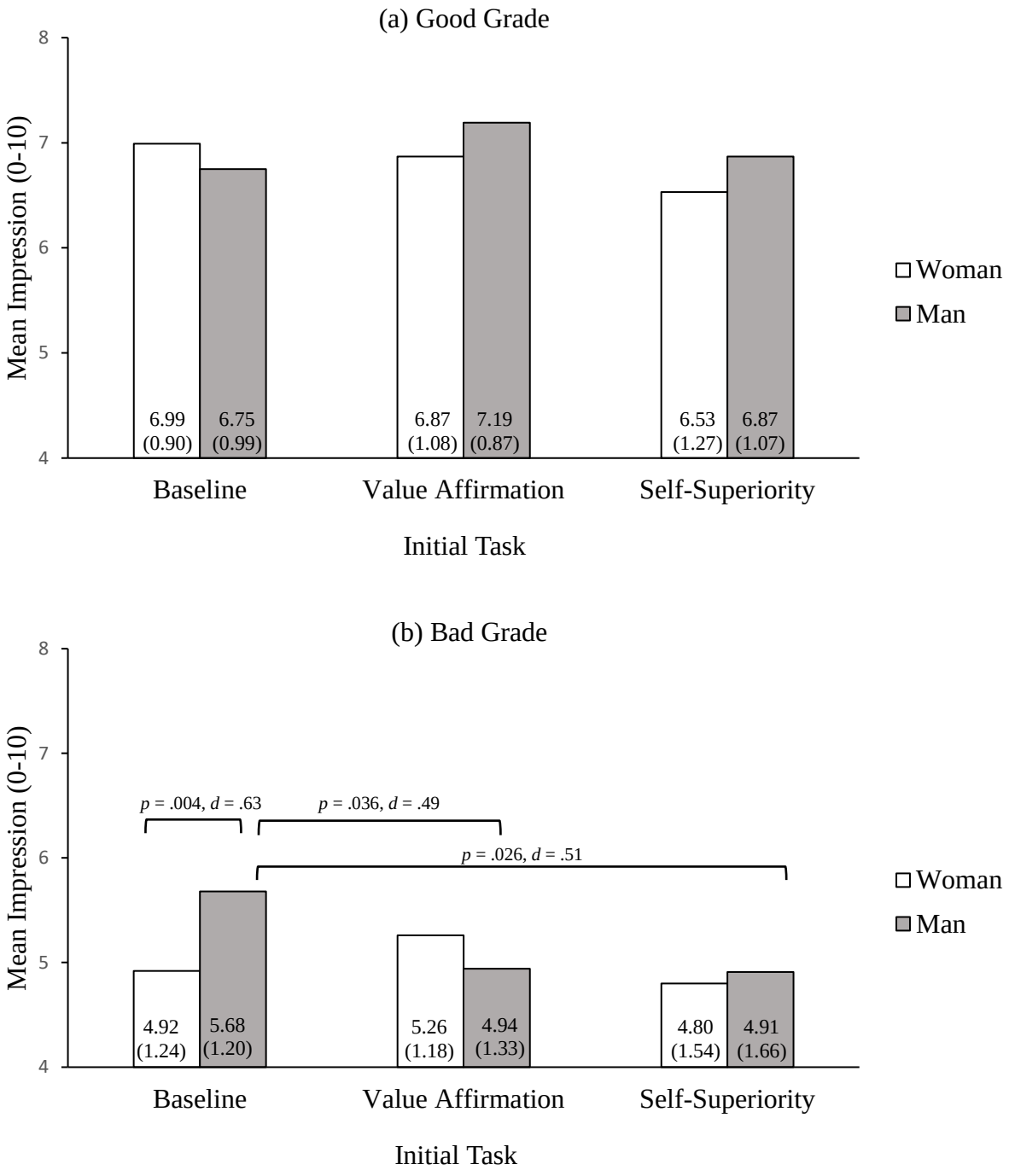

Figure 2. The interactive effect of professor gender and initial task on impressions in (panel a) the good-grade condition and (panel b) the bad-grade condition (*SD*s between brackets).

Online supplement for Hoorens, V., Dekkers, G., and Deschrijver, E. (2020). Gender bias in student evaluations of teaching: Students' self-affirmation reduces the bias by lowering evaluations of male professors. *Sex Roles*. Vera Hoorens, KU Leuven, Belgium. Email: Vera.Hoorens@kuleuven.be

ANALYSES EXCLUDING PARTICIPANTS WHO HAVE NOT SUCCESSFULLY SELF-AFFIRMED

Manipulation Checks

We checked if the grade manipulation was successful by analyzing participants' satisfaction scores, participants' general impressions of the grading system, and participants' impressions as derived from the grading system scale. Besides grade, we included the independent variables initial task and professor to allow us to assess if the grade manipulation was equally successful across the conditions of the other independent variables. Each ANOVA thus included grade (good, bad), professor (man, woman), and initial task (value-affirmation, self-superiority, baseline) as between-subjects variables.

All analyses showed a main effect of grade. Participants were more satisfied with a good grade (M = 8.24; SD = 1.52) than with a bad grade (M = 1.97; SD = 1.54), $F(1,535) = 2276.22$, $p < .001$, $\eta^2_{part} = 0.81$. In addition, general impressions of the grading system were higher in the good-grade condition (M = 6.81; SD = 1.65) than in the bad-grade condition (M = 4.33; SD = 1.54), $F(1,552) = 335.92$, $p < .001$, $\eta^2_{part} = 0.38$, and specific judgments of the grading system were also higher in the good-grade condition (M = 4.89; SD = 0.87) than in the bad-grade condition (M = 4.08; SD = 0.92), $F(1,554) = 116.09$, $p < .001$, $\eta^2_{part} = 0.18$.

Of all participants, 502 (89%) wrote at least one comment. Of these, 456 (81%) wrote about the characteristics of a good professor; 330 (59%) about the format of the course or the exam, and 335 (60%) about elements of the evaluation form. An ANOVA on the number of words written

revealed no effects of our manipulations; all Fs $\leq$ 2.08, ps $\geq$ .150. It seems, then, that participants were generally motivated to fill out the questionnaire as well as they could.

Tests of Hypotheses

As in Sinclair and Kunda (2000, Study 1), participants' impressions of the professor and the course were correlated, r = 0.52; p < .001. We averaged them and report analyses on the average scores. Separate analyses revealed identical patterns.

Hypothesis 1. We predicted that students in the bad-grade condition (but not those in the good-grade condition) would give higher SETs to the male professor than to the female professor. To test this two-way interaction of grade and professor, we subjected the impression scores in the baseline condition to an ANOVA with grade (good, bad) and professor (man, woman) as between-subjects variables.

We found a main effect of grade, $F(1,181) = 96.74$, $p < .001$, $\eta^2_{part} = 0.35$. Participants gave higher SETs after a good grade (M = 6.87; SD = 0.95) than after a bad grade (M = 5.32; SD = 1.27). The main effect of professor was not significant, $F(1,181) = 2.70$, p = .102, $\eta^2_{part} = 0.02$.

In support of Hypothesis 1, grade interacted with professor, $F(1,181) = 9.87$, $p = .002$, $\eta^2_{part} = 0.05$. The cell means are presented in Figure 1. Participants gave higher SETs in the good-grade condition than in the bad-grade condition, but the difference was larger if the professor was a woman, $t(88) = 9.17$; $p < .001$, d = 1.91, 95% CI [-2.52, -1.62], than if the professor was a man, $t(93) = 4.74$; $p < .001$, 95% CI [-1.52, -0.62], d = 0.98. Participants in the bad-grade condition gave a higher SET if the professor was a man rather than a woman, $t(87)$ = 2.96; $p = .004$, d = 0.63, 95% CI [0.25, 1.28], whereas participants with a good grade did not judge the professors differently, $t(94) = 1.24$; $p = .218$, d = 0.25, 95% CI [-0.62, 0.14]. We thus replicated Sinclair and Kunda (2000, Study 1) by showing a gender bias in the bad-grade

condition but not in the good-grade condition. The stage was set to examine the effect of value-affirmation and self-superiority priming on the gender bias.

Hypothesis 2. We predicted that value-affirmation condition and self-superiority priming would eradicate the difference that occurred in the bad-grade condition between the evaluations of male and female professors. We thus predicted a three-way interaction of grade, professor, and initial task. To test the prediction, we conducted an ANOVA with grade (good, bad), professor (man, woman), and initial task (value-affirmation, self-superiority, baseline) as between-subjects variables.

The main effect of grade remained significant, $F(1,528) = 288.73$, $p < .001$, $\eta^2_{part} = 0.35$. Participants gave higher evaluations in the good-grade condition ($M = 6.86$; $SD = 1.05$) than in the bad-grade condition ($M = 5.09$; $SD = 1.39$). A main effect of initial task also occurred, $F(2,528) = 3.60$, $p = .028$, $\eta^2_{part} = 0.01$, but the pairwise contrasts were not significant, Tukey ps $\geq .06$. The main effect of professor was not significant, $F(1,528) = 2.34$, $p = .127$, $\eta^2_{part} \approx 0.00$, nor was the two-way interaction of grade and professor, $F(1,528) = 0.04$, $p = .837$, $\eta^2_{part} \approx 0.00$.

Importantly, and in support of Hypothesis 2, the predicted three-way interaction of grade, professor, and initial task was significant, $F(2,528) = 5.69$, $p = .004$, $\eta^2_{part} = 0.02$. Our hypothesis predicted an interaction of professor and initial task in the bad-grade condition (but not in the good-grade condition). We therefore broke the three-way interaction down per grade.

In the good-grade condition, the professor by initial task interaction was not significant, $F(1,269) = 2.33$, $p = .099$, $\eta^2_{part} = 0.02$. Nor were the main effects of professor, $F(1,269) = 1.22$, $p = .270$, $\eta^2_{part} < 0.01$, and initial task, $F(1,269) = 2.23$, $p = .110$, $\eta^2_{part} = 0.02$. Participants did not differentiate between the male and the female professor at baseline (see above), in the value

affirmation condition, $t(88) = 1.61$, $p = .112$, $d = 0.22$, 95% CI [-0.08, 0.75], or in the self-superiority condition, $t(87) = 1.29$; $p = .201$, $d = 0.21$, 95% CI [-0.17, 0.82].

In the bad-grade condition, and as predicted, the two-way interaction of professor and initial was significant, $F(1,259) = 3.47$, $p = .032$, $\eta^2_{part} = 0.03$. Participants with a bad grade showed a gender bias at baseline (see above), but not after value affirmation, $t(84) = 1.17$; $p = .245$, $d = 0.21$, 95% CI [-0.86, 0.22], or self-superiority priming, $t(88) = 0.29$; $p = .771$, $d = 0.06$, 95% CI [-0.57, 0.77].

We also explored the three-way interaction by examining how the initial task affected participants' impressions per grade and professor. One-way ANOVAs showed an effect of initial task in the bad-grade/male-professor condition only, $F(2,131) = 4.48$, $p = .013$, $\eta^2_{part} = 0.06$. Participants gave higher SETs at baseline than after value-affirmation, Tukey $p = .034$, $d = 0.49$, or self-superiority priming, Tukey $p = .026$, $d = 0.51$, with the latter conditions not differing from each other, Tukey $p = 0.992$. The initial task did not affect impression scores in any other condition: the bad-grade/female-professor condition, $F(2,128) = 1.35$, $p = 0.264$, $\eta^2_{part} = 0.02$, the good-grade/male-professor condition, $F(2,135) = 2.48$, $p = .087$, $\eta^2_{part} = 0.04$, and the good-grade/female professor: $F(2,134) = 2.09$, $p = .127$, $\eta^2_{part} = 0.03$.

In summary, value-affirmation and self-superiority priming eradicated gender bias in the bad-grade condition. The more egalitarian impressions were due to SETs for the male professor becoming lower rather than to SETs for the female professor becoming higher.

Exploratory Analyses

Cover story. We did not predict any effects on the scales bolstering the cover story (specific judgments and judgments of the evaluation form). Yet, we analyzed them exploratorily. Mean ratings of the course and the professor were correlated, $r = .42$; $p < 0.001$.

We averaged them and subjected the average scores to an ANOVA with the same design as the ANOVA on our dependent variables. We found a main effect of grade, $F(1,551) = 276.72$, $p < .001$, $\eta^2_{part} = 0.33$. Participants gave higher ratings after a good grade ($M = 4.69$; $SD = 0.48$) than after a bad grade ($M = 3.98$; $SD = 0.53$). No other effects were significant, Fs ≤ 2.37, ps $\geq .094$. We also analyzed average ratings on the items about the evaluation form. The ANOVA yielded a main effect of grade, $F(1,547) = 5.27$, $p = .022$, $\eta^2_{part} = 0.01$. Participants found the evaluation form better after a good grade ($M = 5.11$; $SD = 0.83$) than after a bad grade ($M = 4.96$; $SD = 0.74$). We also found a marginally significant main effect of professor, $F(1,547) = 3.79$, $p = .052$, $\eta^2_{part} = 0.01$. Participants found the evaluation form better if the professor was a man ($M = 5.10$; $SD = 0.75$) than if she was a woman ($M = 4.97$; $SD = 0.82$). No other effects were significant, Fs ≤ 2.57, ps ≥ 0.078.

Correlations. At many universities, students fill out SET forms voluntarily. If their willingness to do so is associated with their appreciation of male and female professors, that may in itself be a hidden source of bias. We therefore explored how participants' impressions of the course and the professor were related to their attitudes toward the evaluation form. We did so separately for the baseline condition (where gender bias occurred) and the self-affirmation conditions combined (where no gender bias occurred), and separately for the male-professor and female-professor conditions.

At baseline, participants' attitude toward the evaluation form was positively correlated with their impression of the course and the professor. The correlation was significant for the female professor, $r = .25$, $p = .019$, and not for the male professor, $r = .12$, $p = .234$, but the difference between the correlations was not significant, $z = 0.90$, $p = .368$.

In the self-affirmation conditions, the correlation of participants' attitude toward the evaluation form with their impression of the male professor was stronger than at baseline; $r = .36$, $p < .001$; $z = 1.96$, $p = .050$, whereas the correlation with their impression of the female professor was eradicated; $r = .03$, $p = .669$; $z = 1.67$, $p = .095$. As a result, the correlation between participants' attitude toward the evaluation form and their impression was higher for the male than for the female professor, $z = 3.21$, $p = .001$. Eradicating gender bias in a SET-like evaluation task thus went hand in hand with introducing gender bias in attitudes toward the evaluation task itself.