Scaling Down Inequality:
Rating Scales, Gender Bias, and the Architecture of Evaluation

Lauren A. Rivera
Northwestern University
Email: l-rivera@kellogg.northwestern.edu

András Tilcsik
University of Toronto
Email: andras.tilcsik@rotman.utoronto.ca

Forthcoming, *American Sociological Review*
(Near-final manuscript. February 7, 2019 version)

Citation:
Rivera, L. and A. Tilcsik. 2019. "Scaling Down Inequality: Rating Scales, Gender Bias, and the Architecture of Evaluation." *American Sociological Review* (forthcoming).

Acknowledgments
Both authors contributed equally to the work and their names are listed alphabetically. We are grateful to the Institute for Gender and the Economy at the University of Toronto and the Michael Lee-Chin Family Institute for Corporate Citizenship for their generous support and funding. We wish to thank Anne Bowers, Clayton Childress, Jerry Davis, Laura Doering, Sarah Kaplan, Chris Liu, Sida Liu, Kim Pernell-Gallagher, Sameer Srivastava, the Group of Seven, and seminar participants at McMaster University, New York University, University of Michigan, University of Southern California, and Washington University for helpful feedback on earlier drafts.

Abstract

Quantitative performance ratings are ubiquitous in modern organizations—from businesses to universities—yet there is substantial evidence of bias against women in such ratings. This study examines how gender inequalities in evaluations depend on the design of the tools used to judge merit. Exploiting a quasi-natural experiment at a large North American university, we found that the number of scale points used in faculty teaching evaluations—whether instructors were rated on a scale of 6 versus a scale of 10—significantly affected the size of the gender gap in evaluations in the most male-dominated fields. A survey experiment, which presented all participants with an identical lecture transcript but randomly varied instructor gender and the number of scale points, replicated this finding and suggested that the number of scale points affects the extent to which gender stereotypes of brilliance are expressed in quantitative ratings. These results highlight how seemingly minor technical aspects of performance ratings can have a major effect on the evaluation of men and women. Our findings thus contribute to a growing body of work on organizational practices that reduce workplace inequalities and the sociological literature on how rating systems—rather than being neutral instruments—shape the distribution of rewards in organizations.

Women have made great progress in entering highly skilled occupations over the past half-century, but strong disparities in rates of hiring, promotion, and pay persist (England 2010; Roth 2006). Although a number of factors drive gender inequalities in careers, research shows that gendered evaluations of competence play a critical role (Ridgeway 1997, 2006). On average, people rate male workers as significantly more able, likable, and worthy than female workers, even when their qualifications, performance, and behaviors are identical (for reviews, see Foschi, Lai, and Sigerson 1994; Heilman 2001; Quadlin 2018; Ridgeway 2011).

Such research convincingly demonstrates that biased evaluations play a vital role in maintaining gender inequalities, but scholars have paid little attention to the *architecture of evaluation*—specifically, how the design of tools used to judge merit might temper or exacerbate gender gaps in evaluations. Workplace evaluations are classification systems that require individuals to rate an employee's relative quality in selected domains. Rating systems, however, are not neutral instruments; they are sources of power and engines of inequality that strongly shape how people distribute attention, resources, and rewards (Espeland and Sauder 2016; Espeland and Stevens 1998). Nevertheless, scholars have yet to examine how the structure of particular rating schemes used to evaluate workers may influence observed gender inequalities in a given field.

In this article, we analyze a basic element of rating systems that can affect the degree to which gender bias is reflected in evaluations: the number of response categories on a rating scale. We report results from two complementary studies. First, using a quasi-natural experiment, we analyze how a shift from a 10-point to a 6-point scale affected the evaluations of male and female instructors at a professional school of a large North American university. This setting offers a unique opportunity because the shift in the scale—implemented for reasons unrelated to the gender gap in evaluations—represents a quasi-exogenous shock to the rating system and thus helps reveal the relationship between the number of scale points and the size of the gender gap. In particular, by using instructor-course fixed effects, we can adjust for stable differences in instructor and course quality and examine how the same instructors teaching the same courses fared before and after the scale change—and whether the effect of the new scale was different for men and women. Using this approach, we found that the shift in the rating scale eliminated the gender gap in teaching evaluations in the most male-dominated fields.

Second, we consider the explanation that the scale change—rather than reducing the extent to which gender bias is reflected in evaluations—simply reduces opportunities for finer quality differentiation and thus masks actual gender differences in teaching performance. To address this possibility, we conducted a survey experiment that presented all participants with an identical lecture transcript but randomly varied the gender of the instructor who had ostensibly given the lecture and the number of points on the scale that participants used to rate the instructor's performance. Holding instructor quality constant, this experiment replicates the results of our field study and suggests that the number of scale points affects the degree to which gender stereotypes associating brilliant, exceptional performance with men manifest in numeric performance ratings.

These findings highlight how seemingly minor technical aspects of performance ratings can have a major effect on the evaluation of men and women in the workplace. In addition, our study fills important gaps in sociological knowledge by illuminating concrete

organizational practices that can reduce workplace inequalities (see Dobbin, Schrage, and Kalev 2015; Kalev, Dobbin, and Kelly 2006; Williams 2014) and by showing how the design of evaluation tools affects gender dynamics in organizations.

GENDER INEQUALITIES IN PERFORMANCE EVALUATIONS

Performance evaluations are ubiquitous in contemporary organizations (Castilla 2008). Following the scientific turn in management and the emergence of human resources departments as a bureaucratic form, performance evaluations gained popularity as a means to increase efficiency, standardize comparisons between workers, and reduce bias (Dobbin et al. 2015). In the wake of equal opportunity legislation in employment, structured performance evaluations have also become important symbolic tools that organizations use to signal compliance with federal and state anti-discrimination laws (Dobbin 2009; Edelman 2016). Within the broad category of performance evaluations, numeric ratings are among the most common (Murphy and Cleveland 1995).

Despite their intended purpose as "objective" measures of worker performance, a substantial body of research shows systematic bias in performance evaluations against particular groups of workers, including women. Through numerous laboratory and field studies, scholars have shown that women tend to receive significantly lower performance ratings than men, even when their behaviors or skill levels are identical (for a review, see Heilman 2001). When assigning holistic assessments of overall worker quality, managers not only hold women to higher standards in terms of both competence and warmth relative to men (Biernat, Tocci, and Williams 2012; Foschi 1996; Lyness and Heilman 2006), but they also discount women's skills, giving them less credit for their performance (Castilla 2008; Heilman 1995). Compounding this, top-performing women are significantly less likely than men to be described as exceptional performers, geniuses, "stars," or "superstars" due to gender stereotypes of brilliance (Bian, Leslie, and Cimpian 2017; Leslie et al. 2015; Schmader, Whitehead, and Wysocki 2007). Given that performance evaluations form a major basis for promotion, compensation, and termination decisions in many organizations, these patterns are critical for understanding the persistence of gender inequalities in the workplace.

Nevertheless, there is a dearth of research examining what can be done to *reduce* gender inequalities in workplace evaluations, including performance evaluations (Bohnet, Van Geen, and Bazerman 2015; Dobbin et al. 2015; Ridgeway 2006; Williams 2014). Just as the way people evaluate men and women varies based on the structure of the task at hand (see Ridgeway 1997), we argue that the design of tools used to assess merit—or what we term the architecture of evaluation—can influence the degree to which gender stereotypes are expressed in performance appraisals. In this article, we examine one aspect of evaluative structure that can potentially affect the degree to which gender biases manifest: the type of numeric scale used to rate performance.

QUANTIFICATION AND STRATIFICATION

Research on gender inequalities in workplace evaluations typically takes the numeric scale used to assess workers for granted. Numeric performance evaluations, however, are systems of classification that group workers into different categories of worth based on perceived competence. Classification schemes hold tremendous power to shape people's

interpretations of reality (Espeland and Stevens 1998; Lamont and Molnár 2002). They have far-reaching effects on how people rank individuals, objects, and organizations, and how they distribute valued social and material rewards (for reviews, see Brandtner 2017; Lamont 2012; Sauder 2006; Sauder, Lynn, and Podolny 2012). Quantitative classification systems, such as numeric ratings, are particularly powerful sources of inequality because they provide fast, simple, and seemingly neutral bases on which to differentiate actors (Espeland and Stevens 2008; Posselt 2016; Stevens 2007). Prior research has shown that quantitative ratings and rankings of quality play crucial roles in creating and maintaining systems of stratification in diverse settings, ranging from law schools (Espeland and Sauder 2016) to the California wine industry (Benjamin and Podolny 1999).

But crucially, the *structure* of a classification system matters for patterns of inequality that emerge from it (Gould 2003). The number of levels or categories within a classification scheme seems to be particularly consequential for the distribution of material and symbolic rewards. For example, in a study of all-star ratings of equity analysts at U.S. brokerage firms, Bowers and Prato (2018) found that changes in the number of categories rated in a given year had significant effects on analysts' overall visibility and market impact. Crucially, these changes were random and occurred independently of analyst quality. Bowers and Prato (2018:668) argue that the number of categories is consequential because classification schemes, like rating systems, are ultimately hierarchical status systems: "Audiences attend to status . . . and base decisions on it." Thus, changes in the number of categories in an evaluative scheme can "redesign the boundaries of status competition" by shifting how audiences distribute their time, attention, and resources, and by changing opportunities for recognition.

Evaluative schemes with a greater number of categories, for example, enable raters to capture more subtle differences in perceived quality than those with a smaller number of categories, where a broader range of performances may be lumped into a single category (e.g., A through F letter grading schemes versus pass/fail systems in education). But net of such quality considerations, the number of categories present in a rating system may also matter because numbers are more than just counts; they are cultural objects laden with meaning (Espeland and Stevens 2008). In particular, across numerous settings, the number 10 has strong cultural associations with flawless performance (e.g., "a perfect 10"; see Pennington 2016; Stewart 2013). Consequently, evaluators are generally less likely to assign the highest rating to a person on a 10-point scale than on a 5-point scale (Hui and Triandis 1989). Given common gender stereotypes that associate exceptional or brilliant performance with men more than women (Bian et al. 2017), raters might be particularly hesitant to assign a 10/10—an indicator of perfect, brilliant performance—to women. We investigate this possibility and its implications in a setting where gender inequality has received considerable scholarly and public attention: faculty teaching evaluations.

GENDER INEQUALITIES IN FACULTY TEACHING EVALUATIONS

Faculty teaching evaluations provide a ripe setting in which to study the relationship between numeric scales and gender inequalities. Despite substantial gains over the past several decades, gender inequalities in academia remain substantial (Samble 2008). Women are underrepresented in tenure-line roles relative to their representation in PhD programs in both male- and female-dominated disciplines (Rudd et al. 2008). They are overrepresented among adjuncts and non-tenure-line positions, which tend to offer lower

levels of pay, prestige, and job security (Jacobs and Winslow 2004). In many academic fields, women on the tenure track face significant disparities in promotion and pay relative to men (Aguirre 2000; Barbezat and Hughes 2006; Perna 2006).

Gender biases in teaching evaluations contribute to the persistence of gender inequalities in academic careers. Teaching evaluations are central components of faculty hiring, promotion, and compensation decisions (Baldwin and Blattner 2003; Murray 1984). The most common types of teaching evaluations are those where students rate an instructor's performance on a numeric scale. A plethora of experimental and field studies show biases in favor of male faculty in these ratings. Consistent with gender stereotypes of competence (Ridgeway 2011), men are rated as more skilled and able instructors than women; these effects are robust to course content, student self-selection into classes, student learning, and student grades received (Abel and Meltzer 2007; Arbuckle and Williams 2003; Boring, Ottoboni, and Stark 2016; MacNell, Driscoll, and Hunt 2015; McPherson, Jewell, and Kim 2009; Mengel, Sauermann, and Zölitz 2017; Sidanius and Crane 1989; Wagner, Rieger, and Voorvelt 2016). Biases even spill over into assessments of course materials. For example, students give significantly higher ratings to textbooks and readings assigned by men than those assigned by women, even when all course materials are identical across instructors (Mengel et al. 2017). Likewise, students rate female instructors as taking significantly longer to return feedback than male instructors, even when feedback is provided at the exact same time (MacNell et al. 2015).

Additionally, whereas men are judged more for their subject matter expertise, women are judged more for their interpersonal qualities, such as whether they are nice, friendly, or helpful to students (Bennett 1982; Kierstead, D'Agostino, and Dill 1988; Sidanius and Crane 1989). This is consistent with prescriptive stereotypes asserting that women not only are but also *should be* communal and warm (Eagly and Karau 2002).[1]

Gender biases appear not only in numeric ratings of teacher performance, but also in qualitative comments written to describe male and female faculty. Consistent with research on gendered perceptions of brilliance (Bian et al. 2017; Leslie et al. 2015), students in many fields are more likely to describe male versus female faculty as "exceptional," "excellent," or "the best" (Basow 2000; Storage et al. 2016; see also Boring 2017).

Existing studies convincingly demonstrate that faculty teaching evaluations are prone to gender stereotypes of competence, communality, and brilliance. Research shows that biased evaluations can lead to reduced tenure rates and compensation levels for female faculty (Wagner, Rieger, and Voorvelt 2016; see also Murray 1984). It is less clear, however, what policies and practices can help remedy this problem. Gender stereotypes themselves are extremely resistant to change (Ridgeway 2011), but the research on classification and stratification reviewed above suggests the type of rating scale—particularly the number of categories on a scale—might affect the extent to which gender stereotypes manifest in performance ratings.

Shared cultural understandings about the relative value of particular groups of actors, or *status beliefs* (Berger et al. 1977; Webster and Foschi 1988), strongly influence how people classify others in hierarchies, including in rating systems (Sauder et al. 2012). Status beliefs portraying women as being less worthy than men underlie the gender stereotypes reviewed above. They are critical drivers of the more negative teaching evaluations received by

women faculty as well as gender inequalities in workplaces more broadly (Ridgeway 2011; Ridgeway and Correll 2004). If changes in the number of categories in a classification system can indeed "redesign the boundaries of status competition" by refocusing audiences' attention toward or away from particular beliefs and bases of evaluation as relevant (Bowers and Prato 2018), such changes may also make certain status beliefs, including particular gender stereotypes, more or less salient or impactful. Indeed, other properties of evaluations, such as whether they involve qualitative versus quantitative feedback or elicit subjective versus objective data, have been shown to influence the relative salience and impact of gender stereotypes (for reviews, see Biernat and Fuegen 2001; Biernat et al. 2012). Likewise, the number of categories available to rate instructors may affect the magnitude of the gender gap in evaluations.

Yet, the direction of this effect is unclear. On one hand, a reduction in the number of categories, such as our case of moving from a 10-point to a 6-point scale, could increase the gender gap in evaluations. As noted earlier, prior research has documented common stereotypes that link brilliant, extraordinary performance with men more than women (Bian et al. 2017), a general finding that is borne out in research on faculty evaluations, especially in male-dominated fields (e.g., Basow 2000; Boring 2017; Storage et al. 2016). If, because of these stereotypes, students are reluctant to give women the top rating *regardless of the scale used*, then scales with fewer points might disadvantage women. For example, a student might rate a well-performing female instructor 9 on a 10-point scale but only 4 on a 5-point scale—that is, 90 percent of the maximum rating on the 10-point scale versus only 80 percent of the maximum rating on the 5-point scale.

On the other hand, there are reasons to expect that reducing the number of scale points—from 10 to 6 points in our particular context—will benefit women. This prediction is based on two related factors. First, as noted earlier, rating systems with fewer categories tend to be less sensitive to subtle differences in perceived quality. Given research showing that audiences apply stricter standards to female versus male workers (Biernat and Fuegen 2001; Foschi 1996) and more heavily scrutinize their performance for errors (Brewer 1995; Heilman 1995; Rivera 2015), a narrower scale may give audiences fewer opportunities to translate subtle differences in perceived performance—including those driven by gender stereotypes—into numerical differences in ratings.

Second, the highest rating on a scale with more categories might carry a different meaning than the highest rating on a scale with fewer points. Cognitive psychologists make a distinction between individually held, subjective categories of judgment that exist in the minds of evaluators (e.g., "He's an awesome professor," "I hated his teaching style," and "She's a pretty good teacher") and the numerical response categories that exist on a particular scale (Wyer and Carlston 1979). According to this view, raters attempt to map their subjective assessments onto the categories of the scale they are given, and their interpretations of the scale and the different points on it affect how that mapping plays out (Hui and Triandis 1989). Building on this perspective, there is reason to believe that—in addition to wider scales being more sensitive to perceived quality differences—the meanings individuals attach to specific numbers may serve as cognitive anchors, or implicit standards of quality, that shape raters' likelihood of assigning specific scores to particular groups. In other words, certain numbers may prime audiences to adopt more or less stereotypical performance bars. For example, given the cultural associations between the number 10 and perfection described previously, a 6/6 rating may not signify

exceptional or brilliant performance as strongly and unambiguously as a rating of 10/10. Indeed, research shows that evaluators are generally less likely to assign the highest rating to a person on a 10-point scale than on scales with fewer points (Hui and Triandis 1989), suggesting that a 10/10 is a more exclusive category than, say, a 6/6. In other words, whereas a 10/10 rating might be reserved for what raters see as brilliant or perfect performance, a 6/6 rating might be somewhat less exclusive and could encompass not only flawless performance but also what raters perceive as (merely) very good performance. Thus, students evaluating an instructor whom they perceive to be very good but not necessarily extraordinary might be willing to assign a rating of 6/6 but perhaps not a 10/10.

This difference in the meaning of a top rating on a 10-point versus 6-point scale likely matters for the gender gap in faculty evaluations, as well as performance evaluations more broadly. Given the gender biases in evaluation described previously—whereby audiences over-attune to women's weaknesses and under-attune to their strengths (Biernat and Fuegen 2001; Brewer 1995; Foschi 1996; Heilman 1995)—women are significantly less likely than men to be perceived as top performers, especially in male-dominated settings (Biernat et al. 2012; Schmader et al. 2007; Yoder 1991). If students subjectively view female instructors as very good but not brilliant teachers, which prior research suggests they do (e.g., Storage et al. 2016), they might be *less* reluctant to give them 6/6 ratings than 10/10 ratings. This, in turn, may have important implications, because the gender gap in faculty evaluations is often driven by students' relative reluctance to give female instructors the highest ratings on a scale (Boring 2017).

METHODS

We investigate these issues empirically through two complementary studies. First, we analyze data from a quasi-natural experiment to examine how reducing the number of scale points from 10 to 6 affected the ratings received by male and female faculty at a professional school of a large North American university, especially in male-dominated fields, where stereotypes linking superlative performance with men rather than women tend to be most prevalent. Second, we conducted a survey experiment in which we presented all participants with an identical lecture transcript but randomly varied instructor gender and the number of points on the rating scale used in evaluations. The quasi-natural experiment provides data from the field with a high degree of external validity, and it is a unique opportunity to observe the ratings of a given group of instructors under two distinct rating systems; the survey experiment provides a controlled setting in which we can randomly vary our variables of interest *while holding teaching quality constant*.

Quasi-Natural Experiment in the Field

Measuring the causal effect of different rating scales is difficult. Institutions vary in the number of scale points they use in teaching evaluations and in many other ways. There is a great deal of unobserved heterogeneity in course content, instructor quality, gender relations, and institutional cultures. Thus, cross-sectional comparisons of institutions might conflate the effect of different scales with the effect of unobserved factors. A controlled experiment can eliminate such heterogeneity, but it might raise concerns about external validity.

To overcome these issues, we began with a quasi-natural experiment that occurred when—for reasons unrelated to gender—an institution reduced the number of points on its instructor rating scale. This allows us to examine how the ratings of male and female instructors changed after introduction of the new scale. In particular, using instructor-course fixed effects, we are able to examine how the same instructors teaching the same courses fared before and after the scale change—and whether implications of the new scale were different for women and men.

The scale change took place at a professional school of a large, well-regarded North American research university. Per our confidentiality agreement with this institution, we conceal the university's and the school's identities, and we obscure minor details to protect the identity of the school and its students.

For a number of years, the school asked its students to rate instructors on a 10-point scale. However, the school recently switched to a 6-point scale, following the recommendations of an internal committee. There were two main arguments for this change: reducing the number of scale points would simplify the rating system, and a smaller scale might limit "grade inflation" in teaching evaluations. Committee members believed that, because many students mentally converted ratings on the 10-point scale into percentages, they were reluctant to give instructors ratings below 7. With the 6-point scale, the argument went, students would be less likely to convert the points into percentages and, as a result, the ratings might be less skewed. In making the decision for the new scale, the committee did not consider gender issues.

Our data consist of student ratings of instructors for 29 consecutive terms. The 10-point scale was in use during the first 20 terms; during the remaining nine terms, the school used a 6-point scale. In total, our data include 105,034 student ratings of 369 instructors in 235 courses (and 625 instructor-course combinations).[2] Each course falls into one of eight major subject areas.[3]

Overall, 24.4 percent of the instructors in the sample were female, but this number masked important variation across the eight subject areas. Four areas stood out as particularly male-dominated. In these areas, less than one-fifth of the instructors were women (11.1, 13.1, 13.8, and 15.1 percent). In contrast, the proportion of women in the other four fields was substantially higher (38.8, 30.4, 30.0, and 29.2 percent). The gender composition of these eight fields—both the proportion of female faculty in each field and the relative ranking of fields by the presence of women—was consistent with national-level trends in these disciplines. Given the stark difference in the gender composition of these two sets of fields, as well as prior research suggesting that female faculty face distinct challenges in evaluations in the most male-dominated fields (e.g., Storage et al. 2016), we present some of our results separately for the two sets of fields.

As in most institutions, course evaluations at this school included several items. Given our interest in evaluations of individual performance, we focused primarily on students' ratings of an instructor's teaching performance, rather than ratings of the course itself. Instructor ratings represent a particularly important local metric; they are used in decisions about annual salary increases, promotions, tenure, contract extensions, and teaching awards. As a robustness check, however, we also report findings with course ratings as the dependent variable.

For each student rating in our sample, we have a unique code identifying the instructor and the course, know the instructor's gender, the subject area (as classified in the school's course catalog), and the term and the year. In addition, we have a dummy variable *tenured/tenure-track* that equals 1 if the instructor is tenured or tenure-track faculty, and 0 otherwise (e.g., if the instructor is an adjunct or clinical professor or a non-tenure-track lecturer). We also have data on instructor race and know whether an instructor held a PhD; for those who did, we created a time-variant measure of the number of years since their doctoral graduation. Because of confidentiality protections, we did not have access to qualitative comments on the evaluations or individual-level data on the students who provided the ratings.

Overall, 79.7 percent of instructors held a PhD, and 55 percent were tenured or tenure-track faculty. On average, PhD holders had earned their degree 9.5 years before the beginning of our observation period. Most instructors were white (78.1 percent); 16 percent were Asian, 3.5 percent were black, and 2.4 percent were Hispanic. There were no statistically significant differences between male and female instructors in (1) the proportion of PhD holders, (2) the proportion of tenure and tenure-track faculty, or (3) the proportions of various racial groups.

Among PhD holders, however, there was a significant gender difference ($p < .01$) in the number of years since the PhD. On average, men had graduated 10.5 years prior to the beginning of our observation period, whereas women had graduated 6.5 years before the start of the observation window. To a large extent, this difference reflects the relatively small number of female instructors who had obtained a PhD in relevant fields during the 1970s and 1980s. To ensure our analyses do not conflate gender differences with differences in years of experience since the PhD, we conducted robustness checks and report them after presenting our main results.

In our primary regression models, we use fixed effects to adjust for time-invariant characteristics of courses and instructors. This allows us to assess the implications of the scale change while holding constant all stable aspects of course content and instructor quality. Our inferences, in other words, are based on *within-instructor* and *within-course* changes in ratings. Thus, for example, if the school hired a large number of highly rated female instructors after the scale change, their presence would not bias the results.

Instructor-course fixed effects are especially helpful because they mean the coefficients represent changes within instructor-course combinations. This ensures our results are not driven by, for example, the school giving female instructors easier courses to teach after the scale change or matching them more effectively to courses that fit their skills. At the same time, of course, no empirical design is perfect. After presenting results, we consider a number of threats to the validity of this quasi-natural experiment, including the possibility that female instructors tend to show more improvement in their teaching over time than do male instructors, and the concern that there might have been a general trend toward less gender-biased evaluations regardless of the scale change.

Survey Experiment

Our field data—collected unobtrusively in a natural setting over several years—offer significant advantages, but they are not without limitations. One issue is that these data do not allow us to establish whether the observed gender gap is necessarily and exclusively due to gender bias; skeptics might argue that some or all of an observed gap is due to actual differences in the teaching performance of male and female instructors. Some may claim, for example, that systematic gender differences in teaching effectiveness exist, and that male faculty are overrepresented in the right tail of the performance distribution (see Summers 2005). This argument, in turn, might imply that a rating scale with fewer points could help female instructors, not because it limits the expression of gender bias, but because it lumps together brilliant (male) instructors with merely very good (female) instructors. Our field study cannot conclusively rule out this possibility. Another limitation is that our field data came from a single institution and might reflect idiosyncrasies of the local student population rather than broader patterns.

To address these issues and complement our field data, we conducted a second study: an online survey experiment with students from the same type of degree program from dozens of schools. We presented participants with *identical* excerpts from the transcript of a lecture and randomly assigned either a male or a female name to the instructor who had ostensibly given the lecture. We also randomly varied whether participants were asked to rate the instructor on a 10-point or a 6-point scale. Thus, each participant was assigned to one of four conditions, using a 2 (instructor gender: female versus male) × 2 (rating scale type: 6-point versus 10-point) factorial design.

This experimental design has less external validity than a longitudinal field study with years of data from hundreds of actual courses. Yet, it is a useful complement to our field study because it allows us to sample participants from a range of institutions and randomly vary the focal instructor's (perceived) gender and the rating scale *while holding constant instructor quality*.

Our main prediction was that participants would tend to give higher ratings to an instructor believed to be male rather than female, and that the gender gap in evaluations would be larger under a 10-point rating scale than a 6-point scale. Before conducting the study, we preregistered these predictions as well as the planned sample size, the exclusion criteria, and the intended statistical analyses (https://aspredicted.org/blind.php?x=r6hz6x).

We conducted the experiment as an online survey, using participants drawn from respondent panels maintained by SurveyMonkey Audience and Survey Sampling International. The basic sampling frame included degree-seeking students enrolled in on-campus (i.e., rather than online) programs in the same type of professional school as the one we studied with our field data. To obtain a broad but not overly heterogeneous sample, we restricted this sample to students in top-100 degree programs in the United States, as ranked by *U.S. News and World Report* in 2018.

Participants were selected on the basis of their answers to demographic questions they had provided when initially joining a panel; in addition, our survey included a series of screening questions to ensure all respondents fit the above-described sampling criteria. As compensation, participants received a variety of rewards, including gift cards, cash, entry into prize draws, and donations to a charity of their choice. These incentives were

administered by the survey-sampling firms; the average cost of recruitment was approximately $12 per valid respondent.

On average, participants spent seven minutes on the task. Consistent with our a priori exclusion criteria, we excluded from the sample participants who spent less than 60 seconds reading the transcript, because this would suggest a failure to read the stimulus materials sufficiently carefully. We also excluded respondents who indicated, in response to a direct question at the end of the survey, that they had previously watched the TED talk on which the lecture transcript was based.

The sample consisted of 400 students (66.25 percent male) who fit the sampling frame and were not screened out by our exclusion criteria. Participants were randomly assigned to one of four equal-sized experimental conditions (i.e., $n = 100$ per condition). Participants were from 40 different universities, representing all major regions of the United States (Northeast, Midwest, South, and West), with no more than 29 students from any single institution.

All participants read an identical excerpt from the transcript of a lecture on the social and economic implications of technological change. We chose this topic because it has potentially broad appeal, and both technology and economics are traditionally male-dominated fields. The excerpt was about 1,100 words long and included an illustrative figure from a presentation slide. Ostensibly taken from a university lecture, the transcript was actually adapted from a popular TED talk; we excerpted the sections covering the motivation for the topic, the main argument, and a couple of illustrative examples.

For each participant, the instructor was identified either as Professor John Anderson or Professor Julie Anderson. Anderson is a very common surname in the United States, and the first names John and Julie were suitable for our purposes because they send a clear gender signal but otherwise tend to elicit similar reactions with regard to perceived warmth and competence (Newman et al. 2018).

After reviewing the transcript, participants rated the quality of the instructor and the course on either a 6-point or 10-point scale. To explore the mechanisms underlying any potential differences in ratings, participants were then asked to write down the words that first came to mind when they thought of the instructor's teaching performance. Finally, to understand potential differences in the perception of male and female instructors, we included a set of simple items based on prior research on gender stereotypes of performance (e.g., Correll, Benard, and Paik 2007; Rivera and Tilcsik 2016; Rudman and Glick 2001; Storage et al. 2016), asking respondents to indicate the extent to which they viewed the instructor as brilliant, knowledgeable, nice, helpful, and hardworking on five-point Likert scales (1 = strongly disagree, 5 = strongly agree).

To analyze data from the survey experiment, we first examined whether participants gave higher performance ratings when they believed the instructor was male rather than female. We then tested whether the size of the gender gap in ratings was different under the 6-point and 10-point scales. This approach allowed us to replicate the analysis of our field data in a controlled experimental setting where student raters were presented with identical evidence about the teaching performance of male and female instructors. In addition, to probe the underlying mechanisms, we explored how participants' qualitative responses and

quantitative assessments of specific instructor characteristics—such as brilliance—varied by scale type and perceived instructor gender.

FIELD STUDY RESULTS

Descriptive Overview

The histograms in Figure 1 provide a descriptive overview of the distribution of ratings under the original, 10-point-scale system. In the four subject areas with the highest proportion of female instructors, there do not appear to be major gender differences in the shape of the distribution. The distribution in the most male-dominated subject areas, however, reveals a different picture—one that is consistent with the previously documented reluctance of raters in male-dominated fields to evaluate women as performing at the highest levels of excellence. In these fields, 31.4 percent of the ratings male instructors received were a perfect 10; in contrast, female instructors received the highest score in only 19.5 percent of cases. This difference is significant both statistically ($p < .001$) and substantively: men's ratings were 1.6 times more likely than women's to be a 10. Indeed, for men in these fields, 10 was the most common rating (31.4 percent), followed by 9 (22.2 percent) and then 8 (18.9 percent). For women in the same fields, the most common rating was 8 (23.3 percent), followed by 9 (20.3 percent), and only then 10 (19.5 percent). We see this same basic pattern in each of the four fields; male instructors received 10s significantly more often than women in each case. As a result, the average rating in these fields was half a point lower for women than for men (7.7 versus 8.2, $p < .001$).

<Figure 1 about here>

Figure 2 depicts the distribution of ratings under the 6-point scale. In the four areas with the highest proportion of female instructors—just as under the 10-point system—the highest rating was the most common one, followed by the second highest rating, and then the third highest rating. And, as under the old, 10-point system, the distribution of ratings appears fairly similar for male and female instructors. Most interesting for our purposes, however, is what happened in the male-dominated subject areas, where we saw the starkest differences between men's and women's ratings under the 10-point system. Part B of Figure 2 suggests those differences largely disappeared with introduction of the 6-point scale. Under the new system, there was no substantial gender difference in the frequency of receiving a top score: the frequency of top ratings (i.e., 6s) was 41.2 percent for men and 41.7 percent for women. Likewise, there was no longer a major difference in the average rating of men and women (4.91 and 5.01, respectively).

<Figure 2 about here>

Regression Results

Table 1 presents random-effects models, estimated with generalized least squares. Unlike the fixed-effects models we present below, these models do not adjust for all time-invariant course and instructor heterogeneity. Yet, they are still helpful as a starting point: although the main effect of gender is absorbed by fixed effects, it is visible in random-effects models.

<Table 1 about here>

Models 1, 2, and 3 are linear probability models with a binary dependent variable that equals 1 if a rating was at the highest level on the relevant scale (i.e., 10 on the old scale and 6 on the new scale). Models 4, 5, and 6 mirror these models, but the dependent variable is the rating itself.[4]

Looking across all subject areas, the coefficient on *female* in Model 1 implies that under the 10-point scale, the likelihood of receiving the top score of 10 was 5 percentage points lower for women than for men. In addition, the positive coefficient on *6-point scale* implies that instructors were more likely to receive the top score under the new system than under the old one. This is consistent with our expectation that the bar for giving a 6/6 is lower than for giving a 10/10. What is interesting is that women benefitted more from the new system: looking across all subject areas, the likelihood of receiving the top score increased by 7 percentage points for men and by 10 percentage points for women, thus reducing the gap.

Models 2 and 3 show that this basic pattern is especially pronounced in the most male-dominated fields. In the four least male-dominated subject areas, the gender difference under the 10-point scale (i.e., the coefficient on *female*) was not statistically significant (Model 2), and the additional increase that women experienced in the likelihood of receiving a top score (i.e., the coefficient on the interaction term) was both substantively small (just 1 percentage point) and statistically indistinguishable from zero. In contrast, in the four most male-dominated subject areas (Model 3), women received 10 percentage points fewer top ratings than men under the old regime, and women experienced a much larger boost than men in the likelihood of top ratings when the new system was introduced. Whereas men's likelihood of receiving the top score in these fields increased by 8 percentage points under the new system, Model 3 estimates that women's frequency of receiving the top rating increased by 23 percentage points, thus offsetting the previous gender gap in ratings.

Models 4, 5, and 6 reveal similar patterns. Across fields, the average rating unsurprisingly fell as the school moved from a 10-point scale to a 6-point scale. But interestingly, in the most male-dominated fields, men's average rating fell by 3.45 points, whereas it fell by only 2.75 points among women (Model 6). As a result, although women's average rating lagged behind men's—by nearly half a point—under the 10-point scale, it no longer fell short of men's average under the 6-point scale.

The models in Table 2 examine gender differences in the effect of the new system more rigorously, focusing on the most male-dominated fields. These fixed-effect models adjust for instructor and course heterogeneity, showing how the likelihood of receiving the highest rating (Models 7, 8, and 9) and the rating itself as a continuous variable (Models 10, 11, and 12) changed *within* instructor-course combinations. In these models, the main effect coefficient for gender does not appear because it is absorbed by the fixed effects, but the coefficient of the *6-point scale* × *female* interaction is informative because it shows whether the effect of the scale change varied by instructor gender, net of all time-invariant heterogeneity among courses and instructors. In addition, most models in Table 2 include year fixed effects to adjust for trends that had similar effects across all instructors.

<Table 2 about here>

Model 7 shows that the estimated increase in the likelihood of receiving the highest rating on a scale was higher for women than for men in the most male-dominated fields. With the scale change, the estimated likelihood of receiving a scale's top score increased by 14 percentage points more for women than for men, even with instructor-course and year fixed effects, and even when accounting for the possibility that the effect of the scale change varied by race or tenure-track status.

The instructor-course and year fixed effects help account for several alternative explanations, but one might still be concerned that, over time, students' general willingness to assign the highest rating increased—a trend toward "grade inflation" in ratings even without the scale change—and female instructors might have benefitted more from this trend than their male counterparts. Or, maybe there was a general trend over time toward less gender bias in evaluations, leading to a gradual increase in students' willingness to give women the highest rating—even in the absence of any scale change. An increase in the representation of female students over time, for example, might contribute to such a trend.

To address these concerns, Model 8 includes a linear time trend variable, the percentage of female students in the incoming class in the focal year, and the percentage of female instructors (measured as the percentage of sections taught by women in the focal year), as well as the interaction of each of these variables with the focal instructor's gender. Although there was indeed a general trend toward grade inflation in the ratings—the likelihood of receiving the top score on a given scale is estimated to have increased by one percentage point per year—there was no significant gender difference in the effect of the time trend. And, even after adjusting for the time trend and the representation of female students and faculty, as well as their interaction with the instructor's gender, we find that the effect of the scale change was substantially greater for women than for men.

Another concern might be the possible conflation of gender differences with differences in years of experience since receiving the PhD, given that female PhD holders in our sample tended to have fewer years of post-PhD experience than their male counterparts. Model 9 addresses this issue by adjusting for the number of years since the instructor's PhD graduation and its interaction with *6-point-scale*. Even with these variables in the model, the gender difference in the effect of the new scale remains statistically and substantively significant.

Models 10, 11, and 12 mirror Models 7, 8, and 9, except the dependent variable is the rating itself. Not surprisingly, the average rating fell as the school moved from a 10-point to a 6-point scale. But women's average rating fell by a significantly smaller amount than did men's average rating; as a result, female instructors experienced a nearly half-point ratings boost relative to men (Model 10), allowing them to catch up and erase the gender gap that had existed under the 10-point scale. This effect remains largely robust even when adjusting for the time trend and the proportion of female students and instructors (Model 11) and when accounting for differences in the effect of the new scale by years since PhD, race, and tenure-track status (Model 12).

We present several additional specifications in the online supplement. Table S1 disaggregates the results for mandatory core courses and electives. Table S2 examines the curvilinear relationship between post-PhD experience and instructor ratings by adding a quadratic term. Table S3 explores the interaction of instructor gender and race. Our main results of interest—the interaction of gender and scale type in the most male-dominated fields—are robust to these specifications. Finally, the models in Table S4 use ratings of the course, rather than the instructor, as the dependent variable. Consistent with prior research (Mengel et al. 2017), the gender gap in evaluations is somewhat smaller in course ratings than in instructor ratings, but the findings regarding the interaction between scale type and instructor gender are similar to our main results, which is not surprising given the high correlation between instructor and course ratings ($r = .91$).

Alternative Explanations

The models in Table 2 help address several empirical concerns. Thanks to the instructor-course fixed effects, we adjust for time-invariant heterogeneity among instructors and courses (e.g., stable differences in teaching ability and course content) and make inferences based on changes within particular instructor-course combinations. This approach helps address a range of issues, such as the possibility the school hired better female instructors or gave them easier courses to teach after the scale change. In addition, the year fixed effects account for year-specific trends and shocks that affected all instructors.

Our results also show that the change we observe is not driven by a general linear trend toward higher ratings for women (Models 8 and 11). Additional evidence for this comes from Figures S1 and S2 in the online supplement, which show that female instructors experienced an abrupt and discontinuous—rather than gradual—change in the distribution of top ratings after the scale change.

Another approach to addressing the concern that our results simply reflect a gradual trend toward higher ratings for women is to explore whether our findings hold up even when we look at ratings only during a relatively short period of time before and after the scale change. We reran our regressions to focus on the period two years before and after the scale change—a fairly short time during which a radical transformation of student attitudes is unlikely, particularly given the durability of gender stereotypes (Ridgeway 2011). Our coefficient capturing the interaction between instructor gender and scale change remained similar to what we found when looking at the entire period.

Differential attrition may present another empirical concern. One might suspect, for example, that female instructors are subject to more intense selection pressures than men, such that only the very best female instructors remain in the sample by the end of the observation period, thus causing an apparent increase in women's ratings after the scale change. Thanks to instructor fixed effects, however, our inferences reflect within-instructor changes and therefore are based only on instructors observed both before and after the scale change.

Survivorship and differential selection pressures could pose a more nuanced concern: perhaps only women who are highly capable of *improving* their instructional performance, and hence their ratings, stay in the sample over time. As a result, women who are observed both before and after the scale change would tend to show improvement regardless of the

scale change, and instructor fixed effects might not fully capture this trend because its effect is time-variant (i.e., it provides a boost to women's ratings in later but not earlier years). At the same time, as our earlier models show, our results remain robust even when accounting for the interaction between gender and number of years since PhD graduation as well as the interaction between gender and the linear time trend. Thus, women's improvement of their ratings with time and experience does not seem to explain our findings.

For robustness, however, we also ran regression models in which we restricted the sample to faculty members who had already been tenured at the school at the beginning of our observation period. This sample of instructors had presumably not faced strong selection pressures based on their ability to improve their teaching during our time window, and even in this sample, our results remain similar to those reported earlier.

In addition, we wanted to rule out the possible influence of a Hawthorne effect, whereby our results might be tainted by the ratings of students who had directly experienced the scale change and modified their behavior as a result of believing that changes in their rating behavior might be monitored. We reran our models on a sample that did not include ratings from student cohorts that experienced both scales. In these models, too, our coefficients of interest remain significant and similar in magnitude to those reported in Table 2.

Shifts in Ratings

As the final step in our analysis, we explore how instructors' modal ratings changed after introduction of the 6-point scale. Table 3 provides a simple summary; the basic patterns are identical for male and female instructors. Not surprisingly, instructors with a modal rating of 10 under the old system tend to have a modal rating of 6 under the new system. More interestingly, for instructors whose modal rating was 9 under the 10-point scale, the most likely modal rating under the new system is 6. This suggests the move to the 6-point scale benefitted instructors whose performance was perceived to be *very good but not necessarily brilliant or exceptional* (i.e., instructors who tended to receive 9s under the 10-point scale).

<Table 3 about here>

Instructors whose modal rating used to be an 8 also experienced a boost from the scale change, as they tended to move from the third highest (8/10, i.e., 80 percent) modal rating to the second highest modal rating (5/6, i.e., 83 percent). In fact, about one-fifth of instructors whose modal rating used to be an 8 received 6 as their modal rating under the new system.

These shifts, in turn, had implications for the gender gap in ratings, because in male-dominated fields under the 10-point scale, female instructors' most common ratings were 8s and 9s, whereas men's most common rating was already a 10. As 9s tended to turn into 6s, and 8s into 5s (and into 6s in some cases), female instructors in male-dominated fields benefitted from the new scale. Male instructors—whose modal rating was already at the highest possible level under the old system (10/10)—saw less of a boost from the scale change.

As an additional analysis, we examined which female instructors tended to benefit from the scale change. To do so, we explored—in the most male-dominated fields—the ratings of female instructors whose modal rating was *not* in the highest category before the scale change but *was* in the top category after the change. In other words, we identified female instructors who did not typically receive 10/10 ratings under the old system but received 6/6 as their modal rating under the new system. Under the 10-point scale, the median rating of such instructors was 8; their 25th percentile rating was 7, and their 75th percentile rating was 9. Their most frequent rating was 8, followed by 9. This, too, suggests women perceived to be very good (but not extraordinary) instructors—those who typically received 8s and 9s under the old system—benefitted from the scale change and contributed to the erasure of the gender gap in the highest ratings.

SURVEY EXPERIMENT RESULTS

The survey experiment, which held constant instructor quality across all participants and used a sample of students from a broader set of institutions, revealed patterns similar to the field data. Under the 10-point scale, the same instructor received a mean rating of 7.8 (SD = 1.7) when perceived to be male and a mean rating of 7.1 (SD = 2.2) when perceived to be female, a statistically significant gender gap ($p < .05$). When using the 6-point scale, the gap shrank: the instructor received a mean rating of 4.9 (SD = .9) when perceived to be male versus a mean rating of 4.8 (SD = 1.0) when perceived to be female, a difference that is neither statistically nor substantively significant.

The models in Table 4 test whether the difference in the size of the gender gap under the 10-point and 6-point scale is itself statistically significant. In Model 13, the coefficient on *female instructor* indicates a significant gender gap: under the 10-point scale, the perceived female instructor received .64 points lower ratings than the perceived male instructor. In addition, the coefficient on *6-point scale* and the interaction term imply that the average rating of the perceived male instructor was 2.89 points lower under the 6-point scale than under the 10-point scale, whereas the average rating of the perceived female instructor was only 2.28 points (i.e., 2.89 – .61 points) lower under the 6-point scale than under the 10-point scale. As a result, although there was a gender gap of .64 points under the 10-point scale, it largely disappeared under the 6-point scale.

<Table 4 about here>

The gender composition of participants was not equal across conditions; it ranged from 61 percent male participants in the 10-point scale with male instructor condition to 71 percent male participants in the 6-point scale with male instructor condition. Thus, we adjust for participant gender in Model 14 and both participant gender and its interaction with the instructor's perceived gender in Model 15. The main results remain largely the same: consistent with our results from the field, the gender gap is significantly smaller under the 6-point scale than the 10-point scale.[5]

Next, we explored the mechanism underlying this effect. We first observed that the rating category on the 10-point scale on which the perceived male instructor and the perceived female instructor differed most strikingly was the highest (10/10) category. The male instructor received a 10/10 rating in 22 percent of cases, whereas the female instructor

received a 10/10 score in only 13 percent of cases. In contrast, under the 6-point scale, the male and female instructors received the top (6/6) rating with nearly equal frequency (25 and 24 percent, respectively). This suggests that, in line with our results from the field, raters were more willing to give the perceived female instructor the top rating on a 6-point scale than on a 10-point scale.

Participants' qualitative answers to the question about words that first came to mind when they thought of the instructor's teaching performance provided additional insights. Once the survey was closed, a research assistant, who was blind to both our research questions and the experimental conditions, coded participants' qualitative responses for the words that Storage and colleagues (2016) identified as commonly used to describe the highest levels of excellence in written teaching evaluations ("brilliant," "genius," "amazing," and "excellent"). In addition, the research assistant read all responses to inductively identify additional words that signified exceptional (rather than merely good) performance (e.g., "exceptional," "phenomenal," "fantastic," "perfect," and "awesome"). Whereas 12.5 percent of respondents used such words to describe the performance of the perceived male instructor, only 6 percent of respondents described the perceived female instructor in these superlative terms, a statistically significant difference ($p < .05$).[6]

The size of this gender gap did not vary significantly by scale type. What did vary, however, were the qualitative connotations of the highest rating on the 6-point versus 10-point scale, that is, the meanings associated with a 6/6 versus 10/10 rating. In particular, the threshold for receiving a 10/10 seems to have been substantially higher in terms of perceived brilliance than the threshold for receiving a 6/6. Among participants who gave a 10/10 rating, the majority (54.2 percent) used superlative language to describe the instructor's performance. Among participants who gave a 6/6 rating, only 28.6 percent used such language ($p < .05$). These findings suggest the highest rating on a 6-point scale and the highest rating on a 10-point scale differ in the extent to which they are associated with perceptions of superlative teaching performance. In raters' underlying interpretation of numeric evaluation scales, a 10/10 more clearly connotes perceptions of brilliant, extraordinary performance—the kind of performance that is stereotypically associated with men rather than women (see Bian et al. 2017). As a result, for female instructors—who were more likely to be seen as merely good rather than brilliant—a 10/10 rating was particularly difficult to attain. Conversely, because a 6/6 rating did not signify exceptional or brilliant performance as strongly as a 10/10 rating, women—who were less likely than men to be seen as brilliant teachers—benefitted from being assessed on a 6-point rather than 10-point scale.

A similar pattern emerges from examining our Likert-type item for perceived brilliance. When participants believed the instructor was male, 15.5 percent of participants agreed strongly that the professor was brilliant; when participants believed the instructor was female, only 9.5 percent expressed the same view ($p < .10$). Similarly, the average brilliance score was .13 points higher for the perceived male instructor than for the perceived female instructor (3.60 versus 3.47, $p < .10$).[7] As in the qualitative data, however, perceived brilliance seems to have been much less of a requirement for receiving a 6/6 rating than a 10/10 rating: the average brilliance rating was significantly higher ($p < .001$) when the instructor was rated 10/10 (mean = 4.63) than when the instructor was rated 6/6 (mean = 4.02). In addition, 65.7 percent of participants who rated the instructor 10/10 strongly agreed that the instructor was brilliant, whereas only 28.6 percent of those giving a 6/6

score expressed the same view. The average score on the knowledgeable, nice, helpful, and hardworking items was also higher when the instructor was rated 10/10 than when the instructor was rated 6/6, but none of these other mean differences were statistically significant.

Taken together, the survey experiment findings point to two main conclusions. First, even *when raters received identical evidence of teaching performance*, there was a gender gap in evaluations under the 10-point rating scale. Yet, that gap virtually disappeared when the raters used the 6-point scale. Second, raters saw the female instructor as less brilliant than her otherwise identical male counterpart, but their relatively lower expectations of brilliance for a 6/6 rating meant she was more likely to receive the highest rating on the 6-point than on the 10-point scale.

DISCUSSION

Through two studies focusing on faculty teaching ratings, we demonstrated that the design of tools used to judge merit—in this case, the specific numeric scale chosen to assess performance—can powerfully affect gender inequalities in workplace evaluations. Our analysis of a quasi-natural experiment shows that a seemingly minor shift from a 10-point to a 6-point scale helped eliminate previously wide gender gaps in performance evaluations in the most male-dominated fields at a professional school of a large university. Drawing from a complementary survey experiment, we show that this effect is not due to gender differences in instructor quality. Rather, it is driven by differences in the cultural meanings and stereotypes raters attach to specific numeric scales. Whereas the top score on a 10-point scale elicited images of exceptional or perfect performance—and, as a result, activated gender stereotypes of brilliance manifest in raters' hesitation to assign women top scores—the top score on the 6-point scale did not carry such strong performance expectations. Under the 6-point system, evaluators recognized a wider variety of performances—and, critically, performers—as meriting top marks. Consequently, our results show that the structure of rating systems can shape the evaluation of women's and men's relative performance and alter the magnitude of gender inequalities in organizations.

Nevertheless, it is important to emphasize that, although the 6-point scale eliminated the gender *gap* in performance evaluations in both our survey experiment and our field study, we do not argue that it eliminated gender bias. Indeed, in our survey experiment, the gender gap in the frequency of superlative terms describing instructors was not significantly smaller in the 6-point condition than in the 10-point condition, suggesting that even under the 6-point system, differences in the underlying qualitative perceptions of male and female instructors may persist. What we argue is that the architecture of evaluation affects the extent to which gender bias is *reflected* in performance ratings. In our study, the shift from a 10-point to a 6-point scale decreased the gender gap in teaching evaluations by reducing the numerical expression of stereotypes of brilliance in quantitative ratings of teaching performance.

Some scholars may be concerned that this effect is driven not by the relative manifestation and impact of gender stereotypes of brilliance in each rating system, but rather by real gender differences in performance and reduced opportunities to differentiate quality in the 6-point regime. In this view, the 6-point system might lead raters to lump together truly brilliant performance with what is, objectively, merely very good. If this were the case,

moving from a 10-point to a 6-point scale could simply mask real, underlying gender differences in performance (see Summers 2005). However, our field data show that the choice of scale significantly affects ratings of the exact same instructor teaching the same course. Our survey experiment goes one step further by holding performance constant across instructors and replicating our results in a context where arguments that posit actual performance differences by gender do not apply. Thus, the 10-point scale does not better draw out real gender differences in quality. Instead, it contributes to a substantial gender gap based on stereotypes of brilliance that can harm women's opportunities for hiring, promotion, and compensation (Baldwin and Blattner 2003; Murray 1984; Wagner et al. 2016).

Implications for Research on Workplace Inequalities

The core implication of this research is that the architecture of evaluation can reduce or exacerbate gender inequalities in a given field. Research on gendered evaluations and workplace inequalities in sociology, psychology, and organizational behavior has largely neglected the issue of evaluation design. Organizations use a wide variety of tools and metrics to judge worker quality, but a common implicit assumption is that differences between these tools do not affect patterns of inequality, such as the gender gap in performance evaluations. Research in psychometrics centers on evaluative design, measurement, and scale construction, but gender is largely absent from this vast literature; discussions instead focus on the reliability or validity of particular instruments (Furr and Bacharach 2014; Kline 2013; Rust and Golombok 2009).

However, our study shows that evaluative tools are not neutral instruments: their precise design—even factors as seemingly small as the number of categories available in a performance rating system—can have major effects on how female and male workers are evaluated. We focused on performance evaluations, but these findings might have implications for other types of workplace evaluations, including hiring, promotion, and compensation assessments. This research also points to opportunities for future work to examine how other elements of evaluation architecture can affect inequalities in workplace evaluation by gender and other status characteristics, such as race, sexual orientation, disability status, and parenthood, in performance appraisals and beyond. Examples of such evaluative elements include the ordering or wording of question prompts, and whether evaluations are completed anonymously or identifiably, on or offline, or individually versus in groups.

More broadly, our research helps inform debates about the value of bureaucratic personnel management systems. Rationalization and quantification are often presented as solutions that can substantially reduce gender or racial inequalities in workplace evaluations, but existing research shows that in certain cases, implementing formalized procedures can actually exacerbate inequalities (e.g., Dana, Dawes, and Peterson 2013; Jencks 1998). Our study helps explain such contradictory findings by showing that merely looking at the adoption of formal performance metrics is insufficient for understanding the relationship between workplace practices and inequalities: the specific design of the tools used to judge merit also matters.

The occupational context, of course, might be an important scope condition for this effect. We evaluated only one type of scale change in a single industry—higher

education—and cannot conclusively generalize to other occupations. Yet, we believe our results may extend to workplace evaluations in other male-dominated fields where, as in our setting, raters assess performance subjectively and directly observe worker behavior.

Another potential scope condition concerns the specific nature of the rating scale used in evaluations. We studied the effect of a change in scale from a 10-point system, a type of rating scheme that we argue is particularly susceptible to gender stereotypes of exceptionality and brilliance, to a scale that has less robust cultural associations. In addition, 10-point scales may be particularly prone to gender stereotypes, given the link between notions of a "perfect 10" and idealized (and often sexualized) standards of beauty (Pennington 2016; Stewart 2013). Therefore, we cannot conclude that scales with more or fewer scale points, in general, will necessarily increase or decrease the gender gap. Similarly, we do not argue that 6-point scales are ideal rating systems from an equity standpoint. Rather, for those seeking to minimize gender gaps in evaluations, our message is simple: think carefully about the potential benefits and drawbacks of evaluative protocols—and run experiments to test them—because the design of rating systems matters.

Implications for Research on Classification and Stratification

A robust literature examines how classification schemes influence the distribution of resources and rewards in product and financial markets (e.g., Benjamin and Podolny 1999; Bowers and Prato 2018; Espeland and Sauder 2016). Our study extends such research by showing that rating systems are also important drivers of workplace inequalities. We find that the structure of numeric ratings used in performance evaluations affects appraisals of employees' relative value. Given that performance ratings are often tied to important rewards, such as salaries, bonuses, and promotions, rating systems can have direct implications for employees' career trajectories. In addition, this study illuminates how cultural meanings attached to specific numbers can serve as sources of workplace inequalities. Different types of rating systems evoke different types of status beliefs and stereotypes, which in turn serve as cognitive anchors and templates of worth that shape people's definitions of merit, their performance expectations for different groups of workers, and their likelihood of assigning particular scores to a given level of performance.

In addition, our findings contribute to research on expectation states by revealing a novel source of gendered expectations. Existing work has focused on the importance of task-orientation, collective orientation, and mixed-sex groups for triggering biased expectations (see Berger et al. 1977; Correll and Ridgeway 2003). We show that, net of such factors, the sheer presence of certain rating systems can likewise trigger (or mute) gendered expectations and evaluations of performance.

Implications for Research on Faculty Diversity

Finally, our study contributes to research on gender biases in academic careers. Systematic biases against women in letters of recommendation (Madera, Hebl, and Martin 2009; Schmader et al. 2007; Trix and Psenka 2003), peer review (Van der Lee and Ellemers 2015; Wennerås and Wold 1997), hiring decisions (Moss-Racusin et al. 2012; Rivera 2017; Steinpreis, Anders, and Ritzke 1999), attribution of intellectual contributions (Sarsons 2017), citations (Knobloch-Westerwick and Glynn 2013; Malinak, Powers, and Walter

2013), and collaboration opportunities (Knobloch-Westerwick, Glynn, and Huge 2013) keep levels of gender diversity among tenure-line faculty low in many fields. An established and growing body of work shows that teaching evaluations also contribute to gender inequalities within the profession (Abel and Meltzer 2007; Arbuckle and Williams 2003; Baldwin and Blattner 2003; Boring et al. 2016; MacNell et al. 2015; McPherson et al. 2009; Mengel et al. 2017; Sidanius and Crane 1989; Wagner et al. 2016).

Our work provides yet another documentation of the gendered nature of teaching assessments, but we extend prior work in several ways. First, unlike most previous studies on faculty evaluations, our survey experiment demonstrates the existence of a gender gap even when experimentally controlling for instructor performance. Second, our field data provide multiple observations of instructors over time, enabling us to adjust for stable individual-level differences to examine how a given instructor teaching a given course fares under different evaluation systems. Third, consistent with recent research in a European context (Boring 2017), we highlight the difference in scores given to top-performing women and men as a particularly important driver of gender gaps in faculty evaluations. Fourth and most importantly, our work moves beyond documenting the problem to suggesting a remedy to *interrupt* patterns of bias (Williams 2014). We show that not all teaching evaluation schemes are created equal when it comes to issues of inequality; some trigger gendered assessments more than others. Other researchers have suggested costly and elaborate solutions to the problem of gender bias in teaching evaluations, such as relying more heavily on teaching portfolios or faculty observers rather than students to grade performance (e.g., Baldwin and Blattner 2003), but our findings suggest a potentially faster, cheaper, and more easily implementable fix: switch the scale.

Indeed, any organization that regularly collects performance ratings data could run its own experiment by trying out different scales in different time periods—or by randomly assigning different scales to raters in a given period—and then examining the effect on the gender gap. Many organizations already monitor gender differences in performance ratings, and in many cases, experimenting with a different scale would require only modest effort and resources.

At the same time, although we are hopeful about the potential of scale changes to reduce gender gaps in performance evaluations, we believe that guarded optimism is necessary. Prior research shows that gatekeepers strategically shift definitions of merit to protect the privileged status of those in power (Alon 2009; Karabel 2005). Consequently, it is possible that decision-makers in organizations that adopt evaluative design changes may revise definitions of merit if performance evaluations no longer provide advantages for men. For example, in our setting, administrators could view the new teaching ratings as less meaningful differentiators or less reflective of instructor quality. In promotion and compensation decisions, they could attune more to aspects of performance evaluations that still favor men (e.g., qualitative comments, see Biernat et al. 2012); de-emphasize performance evaluations (e.g., weigh course enrollment more heavily); or devise new ways of measuring performance that confirm stereotypes of male superiority. In summary, when it comes to the architecture of evaluation, what matters is not just the rating scale but how decision-makers perceive and use the numbers it produces.

Notes

1. Studies find variation in whether there are *student* gender differences in this effect; the field of study, course content, and country in which the research took place seem to matter (Basow 1995; Boring et al. 2016; Mengel et al. 2017).

2. Because of confidentiality protections, we did not have access to students' qualitative comments. Our survey experiment, however, sheds some light on the meanings associated with 6-point versus 10-point scales.

3. We do not reveal the names and details of the eight fields studied because doing so could potentially reveal the identity of the school, thus violating our confidentiality agreement with the institution.

4. We present generalized least squares models for ease of interpretation and because interaction terms in nonlinear models might produce misleading results (Ai and Norton 2003). For robustness, however, we reran Models 1, 2, and 3 as probit models and Models 4, 5, and 6 as ordinal logit models, and we obtained substantively similar results. In the probit model focused on the most male-dominated fields (i.e., a probit model analogous to Model 3), there was a significant negative coefficient on *female* (b = –.23, SE = .03) and a significant positive coefficient on *6-point scale* (b =.31, SE =.01) and *6-point scale* × *female* (b =.35, SE =.05), indicating the existence of a gender gap in top rating under the 10-point scale and the reduction of that gap under the 6-point scale. Likewise, in the ordinal logit model focused on the most male-dominated fields (i.e., an ordinal logit model analogous to Model 6), there was a significant negative coefficient on *female* (b = –.35, SE = .03) and *6-point scale* (b = –3.51, SE =.03) and a significant positive coefficient on *6-point scale* × *female* (b =.55, SE =.06), pointing to the same conclusion as our linear probability models.

5. The ratings of the instructor and the ratings of the course are highly correlated (r = .82) and show similar patterns. Under the 10-point scale, the course received higher ratings by .52 points (SE = .28, $p < .10$) when taught by the male instructor rather than the female instructor. In contrast, under the 6-point scale, the course received higher ratings by .02 points (SE = .14) when taught by the female instructor, and the gender gap in course ratings was not statistically significant.

6. In contrast, there was no significant difference between the male and female instructor condition in the proportion of qualitative responses that conveyed positive but moderate praise—words describing good rather than brilliant performance—such as "competent," "good," "clear," "interesting," "well informed," and "intelligent."

7. There were no statistically significant differences by instructor gender in the extent to which the instructor was perceived as nice, helpful, or hardworking, but—consistent with prior research (Boring 2017)—the male instructor was rated as somewhat more knowledgeable than the female instructor (4.24 versus 4.12, $p < .10$).

References

Abel, Millicent H., and Andrea L. Meltzer. 2007. "Student Ratings of a Male and Female Professors' Lecture on Sex Discrimination in the Workforce." *Sex Roles* 57(3–4):173–80.

Aguirre, Adalberto, Jr. 2000. *Women and Minority Faculty in the Academic Workplace*. San Francisco, CA: Jossey-Bass.

Ai, Chunrong, and Edward C. Norton. 2003. "Interaction Terms in Logit and Probit Models." *Economics Letters* 80(1):123–29.

Alon, Sigal. 2009. "The Evolution of Class Inequality in Higher Education: Competition, Exclusion, and Adaptation." *American Sociological Review* 74(5):731–55.

Arbuckle, Julianne, and Benne D. Williams. 2003. "Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations." *Sex Roles* 49(9/10):507–16.

Baldwin, Tamara, and Nancy Blattner. 2003. "Guarding Against Potential Bias in Student Evaluations: What Every Faculty Member Needs to Know." *College Teaching* 51(1):27–32.

Barbezat, Debra A., and James H. Hughes. 2006. "Salary Structure Effects and the Gender Pay Gap in Academia." *Research in Higher Education* 46(6):621–40.

Basow, Susan. A. 1995. "Student Evaluations of College Professors: When Gender Matters." *Journal of Educational Psychology* 87(4):656-665.

Basow, Susan A. 2000. "Best and Worst Professors: Gender Patterns in Students' Choices." *Sex Roles* 43(5/6):407–17.

Benjamin, Beth A., and Joel M. Podolny. 1999. "Status, Quality, and Social Order in the California Wine Industry." *Administrative Science Quarterly* 44(3):563–89.

Bennett, Sheila. 1982. "Student Perceptions of and Expectations for Male and Female Instructors: Evidence Relating to the Question of Gender Bias in Teaching Evaluation." *Journal of Educational Psychology* 74(2):170–9.

Berger, Joseph, M. Hamit Fişek, Robert Z. Norman, and Morris Zelditch Jr. 1977. *Status Characteristics and Social Interaction: An Expectation States Approach*. New York: Elsevier.

Bian, Lin, Sarah-Jane Leslie, and Andrei Cimpian. 2017. "Gender Stereotypes about Intellectual Ability Emerge Early and Influence Children's Interests." *Science* 355(6323):389–39.

Biernat, Monica, and Kathleen Fuegen. 2001. "Shifting Standards and the Evaluation of Competence: Complexity in Gender-Based Judgment and Decision Making." *Journal of Social Issues* 57(4):707–24.

Biernat, Monica, M. J. Tocci, and Joan C. Williams. 2012. "The Language of Performance Evaluations: Gender-Based Shifts in Content and Consistency of Judgment." *Social Psychological and Personality Science* 3(2):186–92.

Bohnet, Iris, Alexandra van Geen, and Max Bazerman. 2015. "When Performance Trumps Gender Bias: Joint vs. Separate Evaluation." *Management Science* 62(5):1225–34.

Boring, Anne. 2017. "Gender Biases in Student Evaluations of Teaching." *Journal of Public Economics* 145:27–41.

Boring, Anne, Kellie Ottoboni, and Philip B. Stark. 2016. "Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness." ScienceOpen Research (doi: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1).

Bowers, Anne, and Matteo Prato. 2018. "The Structural Origins of Unearned Status: How Arbitrary Changes in Categories Affect Status Position and Market Impact." *Administrative Science Quarterly* 63(3):668–99.

Brandtner, Christof. 2017. "Putting the World in Orders: Plurality in Organizational Evaluation." *Sociological Theory* 35(5):200–227.

Brewer, Marilyn. 1995. "In-Group Favoritism: The Subtle Side of Intergroup Discrimination." Pp. 101–17 in *Behavioral Research and Business Ethics*, edited by D. Messick and A. Tenbrunsel. New York: Russell Sage.

Castilla, Emilio J. 2008. "Gender, Race, and Meritocracy in Organizational Careers." *American Journal of Sociology* 113(6):1479–526.

Correll, Shelley J., Stephen Benard, and In Paik. 2007. "Getting a Job: Is There a Motherhood Penalty?" *American Journal of Sociology* 112(5):1297–339.

Correll, Shelley J., and Cecilia L. Ridgeway. 2003. "Expectation States Theory." Pp. 29–51 in *The Handbook of Social Psychology*, edited by J. Delamater. New York: Kluwer Academic Press.

Dana, Jason, Robyn Dawes, and Nathanial Peterson. 2013. "Belief in the Unstructured Interview: The Persistence of an Illusion." *Judgement and Decision Making* 8(5):512-20.

Dobbin, Frank. 2009. *Inventing Equal Opportunity*. Princeton, NJ: Princeton University Press.

Dobbin, Frank, Daniel Schrage, and Alexandra Kalev. 2015. "Rage against the Iron Cage: The Varied Effects of Bureaucratic Personnel Reforms on Diversity." *American Sociological Review* 80(5):1014–44.

Eagly, Alice H., and Steven J. Karau. 2002. "Role Congruity Theory of Prejudice toward Female Leaders." *Psychological Review* 109(3):573–98.

Edelman, Lauren B. 2016. *Working Law: Courts, Corporations, and Symbolic Civil Rights*. Chicago: University of Chicago Press.

England, Paula. 2010. "The Gender Revolution: Uneven and Stalled." *Gender & Society* 24(2):149–66.

Espeland, Wendy N., and Michael Sauder. 2016. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. New York: Russell Sage.

Espeland, Wendy N., and Mitchell L. Stevens. 1998. "Commensuration as a Social Process." *Annual Review of Sociology* 24:313–43.

Espeland, Wendy N., and Mitchell L. Stevens. 2008. "A Sociology of Quantification." *European Journal of Sociology* 49(3):401–36.

Foschi, Martha. 1996. "Double Standards in the Evaluation of Men and Women." *Social Psychology Quarterly* 59(3):237–54.

Foschi, Martha, Larissa Lai, and Kirsten Sigerson. 1994. "Gender and Double Standards in the Assessment of Job Applicants." *Social Psychology Quarterly* 54(4):326–39.

Furr, R. Michael, and Verne R. Bacharach. 2014. *Psychometrics: An Introduction*. Thousand Oaks, CA: Sage.

Gould, Roger. 2003. *Collusion of Wills: How Ambiguity About Social Rank Breeds Conflict*. Chicago: University of Chicago Press.

Heilman, Madeline E. 1995. "Sex Stereotypes and Their Effects in the Workplace: What We Know and What We Don't Know." *Journal of Social Behavior and Personality* 10(6):3–26.

Heilman, Madeline E. 2001. "Description and Prescription: How Gender Stereotypes Prevent Women's Ascent up the Organizational Ladder." *Journal of Social Issues* 57(4):657–74.

Hui, C. Harry, and Harry C. Triandis. 1989. "Effects of Culture and Response Format on Extreme Response Style." *Journal of Cross-Cultural Psychology* 20(3):296–309.

Jacobs, Jerry A., and Sarah E. Winslow. 2004. "Overworked Faculty: Jobs, Stresses, and Family Divides." *Annals of the American Academy of Political and Social Science* 596(1):104–29.

Jencks, Christopher, and Meredith Phillips, eds., 1998. *The Black-White Test Score Gap*. Washington, D.C.: Brookings Institution Press.

Kalev, Alexandra, Frank Dobbin, and Erin Kelly. 2006. "Best Practices or Best Guesses? Assessing the Efficacy of Corporate Affirmative Action and Diversity Policies." *American Sociological Review* 71(4):589–617.

Karabel, Jerome. 2005. *The Chosen: The Hidden History of Admission and Exclusion at Harvard, Yale, and Princeton*. Princeton, NJ: Princeton University Press.

Kierstead, Diane, Patti D'Agostino, and Heidi Dill. 1988. "Sex Role Stereotyping of College Professors: Bias in Students' Ratings of Instructors." *Journal of Educational Psychology* 80(3):342–4.

Kline, Paul. 2013. *The New Psychometrics: Science, Psychology and Measurement*. New York: Routledge.

Knobloch-Westerwick, Silvia, and Carroll J. Glynn. 2013. "The Matilda Effect: Role Congruity Effects on Scholarly Communication." *Communication Research* 40(1):3–26.

Knobloch-Westerwick, Silvia, Carroll J. Glynn, and Michael Huge. 2013. "The Matilda Effect in Science Communication: An Experiment on Gender Bias in Publication Quality Perceptions and Collaboration Interest." *Science Communication* 35(5):603–25.

Lamont, Michèle. 2012. "Toward a Comparative Sociology of Valuation and Evaluation." *Annual Review of Sociology* 38:201–21.

Lamont, Michèle, and Virag Molnár. 2002. "The Study of Boundaries in the Social Sciences." *Annual Review of Sociology* 28:167–95.

Leslie, Sarah-Jane, Andrei Cimpian, Meredith Meyer, and Edward F. Freeland. 2015. "Expectations of Brilliance Underlie Gender Distributions across Academic Disciplines." *Science* 16(January):262–5.

Lyness, Karen S., and Madeline E. Heilman. 2006. "When Fit Is Fundamental: Performance Evaluations and Promotions of Upper-Level Female and Male Managers." *Journal of Applied Psychology* 91(4):777–85.

MacNell, Lillian, Andrea Driscoll, and Adam Hunt. 2015. "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching." *Innovative Higher Education* 40(4):291–303.

Madera, Juan M., Michelle R. Hebl, and Randi C. Martin. 2009. "Gender and Letters of Recommendation for Academia: Agentic and Communal Differences." *Journal of Applied Psychology* 94(6):1591–99.

Malinak, Daniel, Ryan Powers, and Barbara F. Walter. 2013, "The Gender Citation Gap in International Relations." *International Organization* 67(4): 889–922.

McPherson, Michael A., R. Todd Jewell, and Myungsup Kim. 2009. "What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes." *Eastern Economic Journal* 35(1):37–51.

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz. 2017. "Gender Bias in Teaching Evaluations." IZA Institute of Labor Economics Working Paper 11000.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 2012. "Science Faculty's Subtle Gender Biases Favor Male Students." *Proceedings of the National Academy of Sciences* 109(41):16474–79.

Murphy, Kevin R., and Jeannette Cleveland. 1995. *Understanding Performance Appraisal*. New York: Sage.

Murray, Harry G. 1984. "The Impact of Formative and Summative Evaluation of Teaching in North American Universities." *Assessment and Evaluation in Higher Education* 9(2):117–32.

Newman, Leonard S., Mingxuan Tan, Tracy L. Caldwell, Kimberley J. Duff, and E. Samuel Winer. 2018. "Name Norms: A Guide to Casting Your Next Experiment." *Personality and Social Psychology Bulletin* 44(10):1435-1448.

Pennington, Ronna. 2016. "Pythagoras from a Different Angle: The Cult Leader Who Influenced Truth, Order, and Beauty." Retrieved January 18, 2018 (https://hubpages.com/religion-philosophy/Pythagoras-from-a-Different-Angle-The-Cult-Leader-Who-Influenced-Truth-Order-and-Beauty).

Perna, Laura W. 2006. "Sex Differences in Faculty Tenure and Promotion: The Contribution of Family Ties." *Research in Higher Education* 46(3):277–30.

Posselt, Julie R. 2016. *Inside Graduate Admissions: Merit, Diversity, and Faculty Gatekeeping*. Cambridge, MA: Harvard University Press.

Quadlin, Natasha. 2018. "The Mark of a Woman's Record: Gender and Academic Performance in Hiring." *American Sociological Review* 83(2):331–60.

Ridgeway, Cecilia L. 1997. "Interaction and the Conservation of Gender Inequality: Considering Employment." *American Sociological Review* 62(2):218–35.

Ridgeway, Cecilia L. 2006. "Gender as an Organizing Force in Social Relations: Implications for the Future of Inequality." Pp. 265–87 in *The Declining Significance of Gender*? edited by F. D. Blau, M. C. Brinton, and D. B. Grusky. New York: Russell Sage.

Ridgeway, Cecilia L. 2011. *Framed by Gender: How Gender Inequality Persists in the Modern World*. New York: Oxford University Press.

Ridgeway, Cecilia L., and Shelley J. Correll. 2004. "Motherhood as a Status Characteristic." *Journal of Social Issues* 60(4):683–700.

Rivera, Lauren A. 2015. *Pedigree: How Elite Students Get Elite Jobs*. Princeton, NJ: Princeton University Press.

Rivera, Lauren A. 2017. "When Two Bodies Are (Not) a Problem: Gender and Relationship Status Discrimination in Academic Hiring." *American Sociological Review* 82(6):1111–38.

Rivera, Lauren A., and András Tilcsik. 2016. "Class Advantage, Commitment Penalty: The Interplay of Social Class and Gender in an Elite Labor Market." *American Sociological Review* 81(6):1097-1131.

Roth, Louise Marie. 2006. *Selling Women Short: Gender and Money on Wall Street*. Princeton, NJ: Princeton University Press.

Rudd, Elizabeth, Emory Morrison, Renate Sadrozinski, Maresi Nerad, and Joseph Cerny. 2008. "Equality and Illusion: Gender and Tenure in Art History Careers." *Journal of Marriage and Family* 70(1):228–38.

Rudman, Laurie A., and Peter Glick. 2001. "Prescriptive Gender Stereotypes and Backlash Toward Agentic Women." *Journal of Social Issues* 57(4):743-62.

Rust, John, and Susan Golombok. 2009. *Modern Psychometrics: The Science of Psychological Assessment*. New York: Routledge.

Samble, Jennifer N. 2008. "Female Faculty: Challenges and Choices in the United States and Beyond." *New Directions for Higher Education* 143:55–62.

Sarsons, Heather. 2017. "Gender Differences in Recognition for Group Work." Working paper, University of Toronto.

Sauder, Michael. 2006. "Third Parties and Status Systems: How the Structures of Status Systems Matter." *Theory & Society* 35(3):299–321.

Sauder, Michael, Freda Lynn, and Joel M. Podolny. 2012. "Status: Insights from Organizational Sociology." *Annual Review of Sociology* 37:267–83.

Schmader, Toni, Jessica Whitehead, and Vicki H. Wysocki. 2007. "A Linguistic Comparison of Letters of Recommendation for Male and Female Chemistry and Biochemistry Job Applicants." *Sex Roles* 57(7–8):509–14.

Sidanius, Jim, and Marie Crane. 1989. "Job Evaluation and Gender: The Case of University Faculty." *Journal of Applied Social Psychology* 19(2):174–97.

Steinpreis, Rhea E., Katie A. Anders, and Dawn Ritzke. 1999. "The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study." *Sex Roles* 41(7–8):509–28.

Stevens, Mitchell L. 2007. *Creating a Class: College Admissions and the Education of Elites*. Cambridge, MA: Harvard University Press.

Stewart, Ian. 2013. "Number Symbolism." *Encyclopaedia Britannica*. Retrieved January 18, 2018 (https://www.britannica.com/topic/number-symbolism#ref248167).

Storage, Daniel, Zachary Horne, Andrei Cimpian, and Sarah-Jane Leslie. 2016. "The Frequency of 'Brilliant' and 'Genius' in Teaching Evaluations Predicts the Representation of Women and African Americans across Fields." *PLOS One* 11(3):e0150194.

Summers, Lawrence H. 2005. "Remarks at NBER Conference on Diversifying the Science & Engineering Workforce" (https://www.harvard.edu/president/speeches/summers_2005/nber.php).

Trix, Frances, and Carolyn Psenka. 2003. "Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty." *Discourse & Society* 14(2):191–220.

Van der Lee, Romy, and Naomi Ellemers. 2015. "Gender Contributes to Personal Research Funding Success in The Netherlands." *Proceedings of the National Academies of Science* 112(40):12349–53.

Wagner, Natascha, Matthias Rieger, and Katherine Voorvelt. 2016. "Gender, Ethnicity, and Teaching Evaluations: Evidence from Mixed Teaching Teams." *Economics of Education Review* 54:79–94.

Webster, Murray, Jr., and Martha Foschi. 1988. *Status Generalization: New Theory and Research.* Stanford, CA: Stanford University Press.

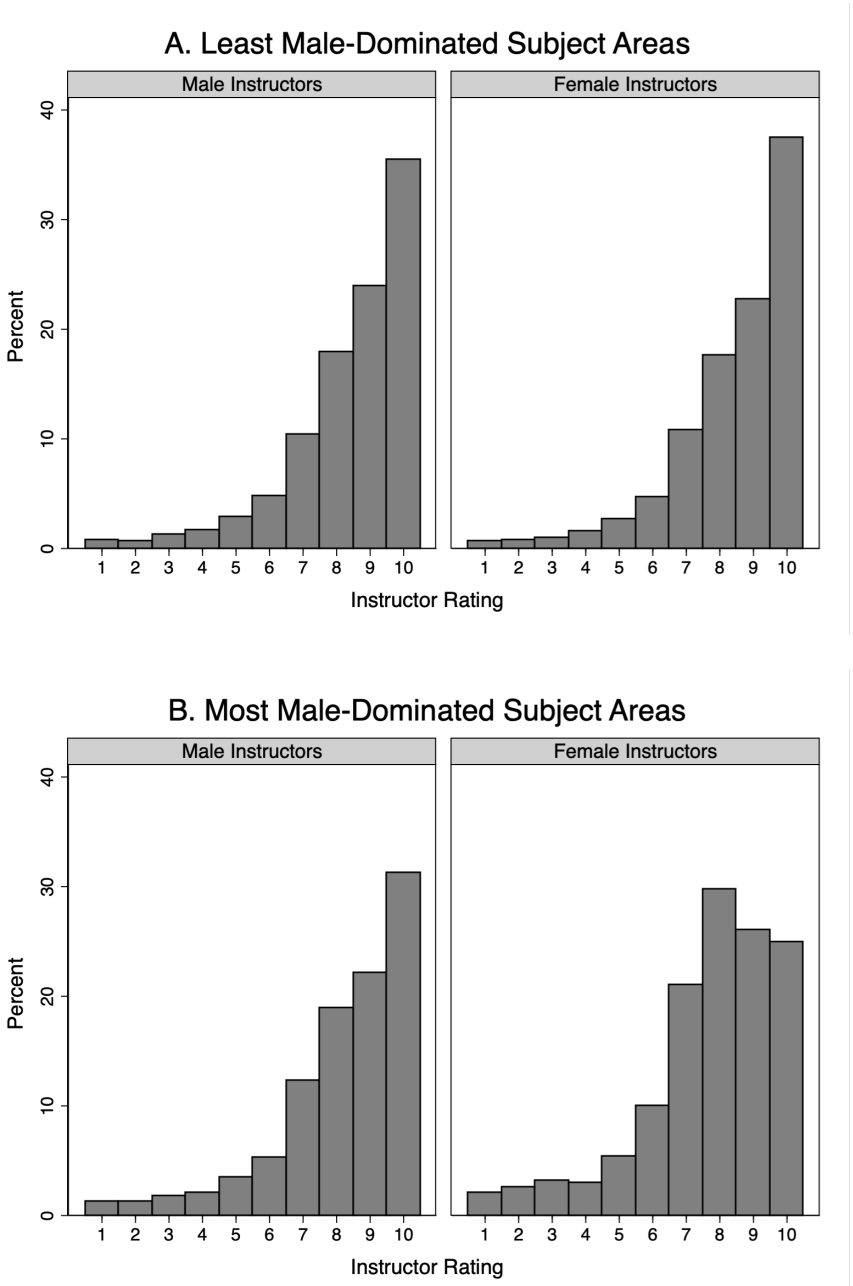
Wennerås, Christine, and Agnes Wold. 1997. "Nepotism and Sexism in Peer-Review." *Nature* 387:341–43.

Williams, Joan C. 2014. "Hacking Tech's Diversity Problem." *Harvard Business Review* 92(10):94–100.

Wyer, Robert S., and Donal E. Carlston. 1979. *Social Cognition, Inference, and Attribution.* Hillsdale, NJ: Lawrence Erlbaum Associates.

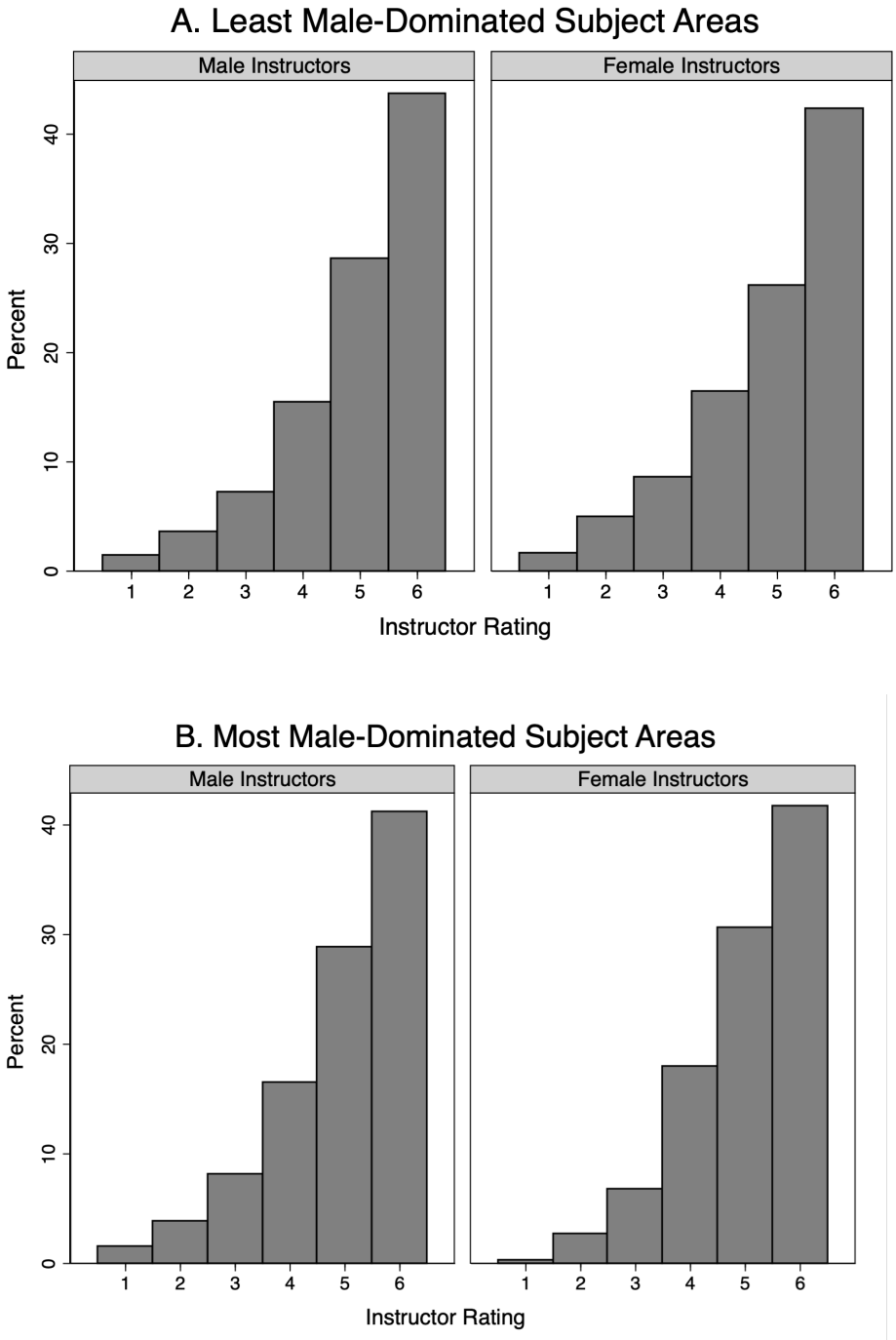
Yoder, Janice D. 1991. "Rethinking Tokenism: Looking Beyond Numbers." *Gender and Society* 5(2):178–92.

Figure 1. Distribution of Ratings under the 10-Point-Scale System



Note: Number of ratings in the least male-dominated subject areas (i.e., the four subject areas with the highest proportion of female instructors) = 29,000 (male instructors) and 12,264 (female instructors). Number of ratings in the most male-dominated subject areas (i.e., the four subject areas with the lowest proportion of female instructors) = 27,674 (male instructors) and 3,272 (female instructors).

Figure 2. Distribution of Ratings under the 6-Point-Scale System



Note: Number of ratings in the least male-dominated subject areas (i.e., the four subject areas with the highest proportion of female instructors) =13,323 (male instructors) and 5,045 (female instructors). Number of ratings in the most male-dominated subject areas (i.e., the four subject areas with the lowest proportion of female instructors) = 13,210 (male instructors) and 1,246 (female instructors).

Table 1. Random-Effects Regressions Predicting Instructor Ratings

	Highest Rating on Scale (0/1)			Rating		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Subject Areas:	All	Least Male-Dominated	Most Male-Dominated	All	Least Male-Dominated	Most Male-Dominated
Female	-.054*	-.031	-.099*	-.207	-.102	-.488*
	(.024)	(.030)	(.044)	(.127)	(.160)	(.246)
6-point scale	.072***	.065***	.080***	-3.525***	-3.603***	-3.447***
	(.004)	(.005)	(.005)	(.013)	(.017)	(.020)
6-point scale × female	.026**	.008	.148***	.112***	.061	.697***
	(.009)	(.010)	(.020)	(.031)	(.033)	(.075)
White	.020	.078*	-.028	.186	.359	.086
	(.025)	(.034)	(.035)	(.131)	(.186)	(.197)
Tenured/tenure-track	-.014	.022	-.064*	.063	.197	-.168
	(.021)	(.026)	(.032)	(.109)	(.144)	(.177)
Constant	.260***	.199***	.316***	7.790***	7.621***	7.920***
	(.026)	(.038)	(.035)	(.137)	(.203)	(.197)
Observations	105,034	59,632	45,402	105,034	59,632	45,402
Wald chi-square	525.0	227.2	368.2	89,374	60,357	31,298

Note: All models are random-effects models, estimated with generalized least squares. Standard errors are in parentheses. The dependent variable in Models 1, 2, and 3 is a dummy that equals 1 if the rating was in the highest category of the relevant scale (10/10 or 6/6). In Models 4, 5, and 6, the dependent variable is the rating itself.

*$p < .05$; **$p < .01$; ***$p < .001$ (two-tailed).

Table 2. Fixed-Effects Regressions Predicting Instructor Ratings in the Most Male-Dominated Fields

	Highest Rating on Scale (0/1)			Rating		
	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
6-point scale	-.026	.018	.017	-3.836***	-3.633***	-3.420***
	(.022)	(.010)	(.019)	(.080)	(.037)	(.070)
6-point scale × female	.136***	.146***	.129***	.461***	.329**	.236*
	(.023)	(.033)	(.032)	(.084)	(.122)	(.116)
6-point scale × white	.022			.043		
	(.016)			(.057)		
6-point scale × tenured/tenure-track	.004			.222***		
	(.015)			(.053)		
Time trend		.011***			.011	
		(.003)			(.011)	
Time trend × female		.011			.085*	
		(.009)			(.033)	
% Female students		.003			.020***	
		(.001)			(.005)	
% Female students × female		-.007			-.012	
		(.005)			(.018)	
% Female instructors		.003			.012*	
		(.002)			(.006)	
% Female instructors × female		.006			.013	
		(.005)			(.018)	
Years since PhD			.027***			.061***
			(.003)			(.012)
6-point scale × years since PhD			-.002*			-.013***
			(.001)			(.002)
Female × years since PhD			-.001			.064*
			(.007)			(.027)
Instructor-course fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	No	Yes	Yes	No	Yes
Constant	.291***	.126*	-.025	8.176***	7.251***	7.362***
	(.008)	(.052)	(.041)	(.031)	(.191)	(.148)
Observations	45,402	45,402	40,131	45,402	45,402	40,131
F	29.2	36.1	31.7	2,023	2,777	1,827

Note: Coefficients for instructor-level variables that are time-invariant (i.e., gender, race, and faculty status) do not appear in this table because they are absorbed by the fixed effects. Standard errors are in parentheses.
*$p < .05$; **$p < .01$; ***$p < .001$ (two-tailed).

Table 3. Changes in Instructors' Modal Ratings

Male Instructors		Female Instructors	
Modal Rating under the 10-Point Scale	Most Likely Modal Rating under the 6-Point Scale	Modal Rating under the 10-Point Scale	Most Likely Modal Rating under the 6-Point Scale
10	6	10	6
9	6	9	6
8	5	8	5
7	4	7	4

Note: This table displays the most likely modal rating of an instructor under the new 6-point scale as a function of the instructor's modal rating under the old 10-point scale; these numbers are identical for male and female instructors.

Table 4. OLS Regressions Predicting Instructor Ratings in the Survey Experiment

	Model 13	Model 14	Model 15
Female instructor	–.640*	–.649*	–.534
	(.277)	(.277)	(.368)
6-point scale	–2.890***	–2.908***	–2.917***
	(.196)	(.196)	(.196)
6-point scale × female instructor	.610*	.627*	.637*
	(.307)	(.307)	(.307)
Male respondent		.184	.275
		(.170)	(.222)
Male respondent × female instructor			–.182
			(.341)
Constant	7.750***	7.638***	7.582***
	(.174)	(.208)	(.231)
F	105.4	80.36	64.46

Note: $N = 400$ (100 per condition). Standard errors are in parentheses.
*$p < .05$; **$p < .01$; ***$p < .001$ (two-tailed).