CAN GENDER AND RACE DYNAMICS IN PERFORMANCE APPRAISALS BE DISRUPTED? THE CASE OF ANCHORING*

Iris Bohnet, Oliver P. Hauser, Ariella Kristal

Abstract:

Performance reviews in firms are common but controversial. Managers' subjective appraisals of their employees' performance and employees' self-evaluations might be affected by demographic characteristics. As self-evaluations are typically shared with managers before they rate employees, social influence via anchoring may also contribute to gender or race differences in performance ratings. Analyzing the data of a multi-national financial services firm, we find gender and intersectionality gaps in self-ratings as well as race and intersectionality gaps in manager ratings, leading to female employees of color being assigned the lowest final ratings. We evaluate a mechanism that could disrupt these dynamics: a quasi-exogenous shock leading to self-evaluations not being shared with managers before they appraised employees. This intervention induced "de-anchoring," leading to lower average manager ratings. However, it did not change overall gender or race dynamics as managers imposed their own gender and race effects (independent of self-evaluations), and also had access to previous years' ratings for most employees. To exclude potential anchors from the past, we focus on employees with no history in the firm during their first year of employment. Among these newcomers, employees who traditionally have given themselves the lowest self-ratings—women of color—benefitted most from the intervention. The intervention did not, however, close the manager-induced race gap overall. These race dynamics were particularly pronounced in the US, negatively affecting Black, Asian and Latinx employees. Counterfactual simulations suggest that 22-28% of Black employees' manager ratings would have to be increased for this race gap to disappear.

JEL Codes: D90, D91, J71, M14

*Bohnet: Harvard Kennedy School, Cambridge, MA, USA. Hauser: Department of Economics, University of Exeter, Exeter, UK. Kristal: Harvard Business School, Cambridge, MA, USA. Correspondence: Iris Bohnet, Harvard Kennedy School, 79 JFK Street, Cambridge, MA 02138; email: iris_bohnet@harvard.edu; phone: +1(617)495-5605. We thank participants in seminars at Carnegie Mellon University, the Center for Economic and Policy Research, Harvard University, and the University of California Berkeley's Science of Diversity and Inclusion Initiative for their helpful comments, and Pivotal Ventures and the UKRI Future Leaders Fellowship for their generous support.

1. Introduction

Annual performance reviews are common. They are part of organizations' efforts towards merit-based performance management systems that measure, incentivize and reward employees' contributions to the organization. In a 2014 survey conducted by the Society for Human Resource Management, 97 percent of organizations reported conducting formal performance appraisals (Society for Human Resource Management, 2014). Often, performance appraisals are directly linked to compensation, promotion, work allocation and termination decisions (Castilla 2008, 2015; Dobbin et al. 2015).

Performance reviews are controversial. Performance is rarely measurable objectively in complex work environments. Subjective performance appraisals that rely on evaluations made by managers are prone to various evaluation biases, including those based on demographic characteristics. In addition to potential manager effects, employees might also contribute to differences in final performance scores through their self-evaluations. In many performance appraisal systems, employees' self-evaluations are shared with managers before managers make up their minds. Thus, final performance scores assigned by managers are likely due to some combination of the true indication of performance, potential effects due to demographic characteristics that affect self- and/or manager assessments, and employees' self-evaluations influencing manager ratings.

This paper examines this potential "social influence" channel by exploring an intervention that could disrupt the link between employee self-ratings and manager assessments. Working with a multi-national financial services firm headquartered in the US, we take advantage of a quasi-exogenous shock to this firm's performance appraisal process. Normally, employee self-evaluations are shared with managers before the latter appraise their employees.

In 2016, without the employees' or managers' prior knowledge, the firm experienced a time crunch due to factors unrelated to the process at hand and was unable to share self-evaluations with managers before they appraised their employees. If managers are typically influenced by their employees' self-ratings, we should see "de-anchoring" in 2016 (we refer to this as the "non-standard year"), with managers' and employees' ratings less correlated with each other in the non-standard year than in standard years where self-evaluations were shared. In addition, we might also see differential effects by employee demographic characteristics: members of groups who rate themselves particularly harshly might benefit from "de-anchoring", while members of groups who rate themselves particularly leniently might be hurt by it.

We focus on the two demographic characteristics available to us, gender and race, and their interactions. Earlier research suggests systematic gender differences (with women giving themselves lower ratings than men) but no race differences in self-ratings in organizational settings (Paustian-Underdahl, Walker and Woehr, 2014). Based on laboratory experiments that show a close link between self-ratings and "supervisor" ratings (Klimoski and Inks, 1990; Shore, Adams and Tashchian, 1998), we therefore expect women to benefit most when self-evaluations are not shared with managers.

Our results can be summarized as follows. In the non-standard year, employees' average performance scores were lower, suggesting that de-anchoring from self-ratings took place when self-ratings were not visible to managers. However, overall gender and race dynamics remained mostly the same. This is not surprising for race as there were no main effects of race on self-ratings. In contrast, we would have expected changes in manager ratings for gender, as we found persistent gender and intersectionality effects in self-ratings, which could have acted as

differential "anchors" for managers in standard years: women, and even more so, women of color, consistently gave themselves lower self-ratings than their respective male counterparts.

Our data suggests two reasons for the persistent race and gender patterns in the non-standard year: First, managers closed (or even reversed) the gender gap in self-ratings in standard years, thus, already addressing what the de-anchoring intervention in the non-standard year could have fixed. Second, in addition, managers could at least in theory have accessed previous years' ratings in all years, including in the non-standard year. While we have no data on how many managers in fact consulted prior ratings, we find some evidence that they might have done so in particular in the non-standard year: the "shadow of the past"—i.e. their own evaluation of the employee in the previous year—was correlated more with their current year's rating of the employee in the non-standard year than in standard years.

In order to rule out the impact of previous years' ratings, which could act as another anchor, we take a closer look at employees during their first year of employment in the firm, the "newcomers," for whom there are no past evaluations. Similar to our full sample, average manager ratings assigned to newcomers were lower in the non-standard than in the standard year, suggesting de-anchoring. The race and gender dynamics still remained mostly unchanged as in the full sample but there was one exception: female employees of color in their first year of employment were evaluated more positively than their male counterparts, on par with White male and female employees, when self-ratings were not available as potential anchors. As our sample size for newcomers is rather small, we interpret this finding only as preliminary evidence that not sharing self-evaluations could benefit those who give themselves the lowest self-evaluations. Notably, however, the intervention had no effect on persistent manager-induced race effects, motivating us to take a closer look at the race dynamics in this firm.

To examine race effects more thoroughly, we focus on the United States where race data is fully available at a relatively granular level. Earlier research for the US has documented consistent race differences in performance evaluations with employees of color being evaluated less favorably than White employees (e.g., Roth et al. 2003; McKay and McDaniel 2006 for meta-analyses). Indeed, our global race findings are heavily driven by the US in particular: the difference between manager and self-ratings was larger for all employee groups of color—including Asian, Black, Latinx and others—than for the White employee group, suggesting persistent manager effects. We employ counterfactual simulations to study the magnitude of these effects in the US. By bootstrapping the US data, we identify how many employees of color would have to receive a higher manager rating for the gap between White employees and Asian, Black and Latinx employees to close. We estimate that between 22% and 28% of Black employees would have to receive a boost in manager ratings for the race gap to close, while fewer than 7-12% of Asian, Latinx and other employees would need to see their manager rating increase by one score (on a 5-point scale).

The remainder of our paper is organized as follows. We first provide a conceptual framework and offer some context about the field site of our study. We then present the impact of the quasi-exogenous shock for all countries, followed by a closer look at the race dynamics in the United States. We conclude with a discussion.

2. Conceptual Framework

Our paper takes as a starting point a performance evaluation process common in many organizations. The task for evaluators (typically managers) is to assess their employees based on their performance in the past year. These standardized evaluation practices began to become

prominent in the 1970s, in response to concerns that managerial discretion in promotion and salary decisions may largely be driven by a combination of arbitrariness, bias or non-performance related characteristics such as seniority (Castilla 2012). Performance appraisals were seen as a means to introduce a common standard for all employees. However, one might suspect that the same problems that prompted the introduction of performance appraisal systems plague them today: evaluators might assess people's performance differently due to taste-based or statistical discrimination based on accurate or biased beliefs about performance differences across groups (Arrow 1973; Bohren et al. 2019; Bohren, Imas and Rosenberg 2019;; Bordalo et al. 2016, 2019; Coffman, Exley and Niederle 2021; Hauser and Bohren, 2021; Phelps 1972). Here we review the evidence on gender and race differences (and, where applicable, discrimination[1]) in the labor market generally as well as for manager and self-ratings, followed by a framework that will guide our empirical analysis.

2.1. Gender

Gender-based discrimination has been studied widely. For example, Goldin and Rouse (2000) showed that female musicians were discriminated against in orchestra auditions when evaluators knew their gender, Moss-Racusin et al. (2012) that science faculty evaluated male applicants for a lab manager positions more highly than otherwise identical female applicants, Bohnet, van Geen and Bazerman (2016) that high-performing women and men were overlooked for counter-stereotypical tasks when evaluated separately, and Quadlin (2018) that employers

[1] As most firms do not collect objective performance measures, most evaluations of employees' performance are subjective. In the absence of objective benchmarks, it is not possible to determine whether gender or race differences are due to true performance differences or the result of discrimination. We therefore avoid the use of the term "discrimination" when talking about gender or race differences in (subjective) evaluations.

used gendered standards for job applicants looking for competence in men and likability in women (for a review, see Bohnet, 2016 and Bertrand and Duflo, 2017).

For self-ratings, gender differences have been documented with women assigning themselves lower ratings than men (Paustian-Underdahl, Walker and Woehr, 2014). Such lower self-ratings could be driven by various factors. For example, under many circumstances women have been shown to be less self-confident (Barber and Odean 2001; Bordalo et al. 2019), less willing to take risks (Croson and Gneezy 2009, Buser, Niederle, and Oosterbeek 2014), less likely to contribute ideas in male-stereotyped domains (Gallus and Heikensten, 2019, 2020), less likely to negotiate (Babcock and Laschever 2003), less likely to compete (Niederle and Vesterlund 2007), less likely to self-promote or ask for a promotion (Bosquet, Combes and García-Peñalosa 2019; Exley and Kessler 2019; Hospido, Laeven and Lamo 2019) and more affected by self-stereotyping than men (Coffman 2014). Many of these behavioral patterns have been described as responses to environments that punish women for counter-stereotypical behaviors, often referred to as "social backlash" (Bowles, Babcock and Lai 2007).

The evidence on gender differences in manager-assessed performance is, however, mixed. A meta-analysis of the impact of gender on performance appraisals found a small gender gap in performance evaluations favoring men but large variations across studies with men, women or neither being evaluated more favorably (Joshi et al. 2015; DeNisi and Murphy 2017 for a review).

2.2. Race

Race-based discrimination has been widely documented in the labor market, the criminal justice system and in many other domains (e.g., Bertrand and Mullainathan 2004; Pager and Pedulla 2015; Rosette, Akinola and Ma 2015; Arnold et al. 2020). For example, Quillian et al.

(2017) document discrimination against Black and Latinx job seekers by analyzing the likelihood that job applicants receive a call back from an employer based on data from all available field experiments since 1989. They find a decline in discrimination against Latinx over time but no change in the levels of discrimination against Black job seekers. In sports, the referees' racial bias affected the likelihood that personal fouls were called in basketball and how pitches were evaluated in baseball (Price and Wolfers 2010; Parsons et al. 2011). While the research on race has mostly focused on the US, a few studies have looked at differences in treatment by ethnic background in other countries as well (for cashiers in grocery stores and biased managers in France, see Glover, Pallais and Pariente 2017).

While there is no body of evidence describing the presence of race differences in self-ratings,[2] race differences have consistently been documented for manager ratings. For instance, Greenhaus, Parasuraman and Wormley (1990) report that Black employees received lower performance ratings and lower promotability scores compared to White employees in three US companies. Castilla (2012) also reports that Black employees (but not Asian American or Latinx employees) received lower manager ratings than White employees. A meta-analysis found that Black-White differences in work performance evaluations are common across many studies and contexts (McKay and McDaniel 2006).

2.3. Intersectionality between gender and race

Studying the intersection of gender and race has gained importance in recent years (Crenshaw 2017). For example, Milkman, Akinola and Chugh (2015) document discrimination in willingness to mentor students in academia with professors preferring White men to all other

[2] Phelan and Rudman (2010) provide some evidence that Black people reduce their public self-appraisal when they perform particularly well on a task, in anticipation of potential social backlash.

possible intersectional groups combined. Some research suggests "double jeopardy" with women of color being disadvantaged by both their gender and their race relative to White men (Beale 1970; Berdahl and More 2006; and for a review, Rosette 2018). For example, Black women who led struggling organizations were evaluated more negatively than Black men or White women (Rosette and Livingston 2012), and Black and Latinx women experienced the largest wage gap relative to White men, resulting from lower likelihood of promotions and salary increases over time (Castilla 2008). Other studies present evidence on "intersectional invisibility" where people with intersecting identities are associated less with their constituent gender or race groups, decreasing the likelihood that evaluators apply gender or racial stereotypes (Purdie-Vaughns and Eibach 2008; Ridgeway and Kricheli-Katz 2013). For example, Black women were found to be penalized less for counter-stereotypical behaviors than Black men (Livingston et al. 2012; Biernat and Sesko 2013). To the best of our knowledge, intersectionality has not been studied in the context of performance appraisals.[3] Given the contradicting predictions the two most prevalent conceptualizations of intersectionality make, we treat this as primarily an empirical question.

2.4. Social influence in evaluations

Many performance appraisal systems share common features (for a recent review of performance appraisals, see Schleicher et al. 2019). One such common feature is the ability of both managers and employees to provide assessments of the employee's performance (which we refer to as "manager ratings" and "self-ratings", respectively). Allowing employees to self-rate and share their own impressions with the manager ahead of the manager rating is in part

[3] Castilla (2012) comes closest by examining performance appraisals simultaneously by gender and race (but not their interaction), finding that women get rated more positively than men by managers and Black employees get rated more negatively than White employees, generally consistent with our own results.

motivated by giving employees more "voice" (Dulebohn, 1995), increasing perceptions of procedural justice (Dulebohn and Ferris, 1999) and encouraging positive reception of feedback (Korsgaard, 1996). But doing so can have unintended consequences. For example, Klimoski and Inks (1990) and Shore, Adams and Tashchian (1998) demonstrate that participants in a laboratory experiment evaluated the work of another participant more negatively (positively) when they received a low (high) self-rating, relative to when no self-rating was shared, demonstrating that self-ratings can act as an "anchor."[4] Indeed, even the managers' own recent experience being evaluated during a performance appraisal can anchor them during subsequent evaluations (Latham et al., 2008), as can an example rating before evaluating an employee's performance (Thorsteinson et al., 2008).

2.5. Framework

Our main interest is in examining the social influence pathway from self-ratings to manager ratings. In our empirical setting, we evaluate a quasi-exogenous shock to self-ratings being shared with managers to enable us to answer the following policy-relevant research question: Can eliminating sharing self-evaluations with managers disrupt a social influence channel where employee self-ratings (and gender or race gaps induced by those ratings) affect manager ratings?

First, based on the above literature, we expect that manager ratings will correlate less with self-ratings when the latter are not shared with managers (which we refer to as a process of "de-anchoring" in the non-standard year). Conversely, when self-ratings are provided to

[4] In particular, in ambiguous situations, experts ranging from real estate agents to legal professionals have been shown to be influenced by (often irrelevant) information provided to them—an "anchor"—before they made an expert judgment (Tversky and Kahneman 1974; for a review, see Furnham and Boo 2011).

managers during standard years, we expect them to influence managers' decisions, leading to self-ratings being more correlated with manager ratings in those years.

When self-ratings are shared with managers, it places importance on the level of the self-rating that is being shared with the manager. Unlike in controlled laboratory experiments (e.g. Klimoski and Inks, 1990; Shore, Adams and Tashchian, 1998), the self-ratings in our setting are not randomly assigned (and thus, cannot be treated as exogenous) but they may vary systematically by demographic characteristics. Indeed, the above literature suggests that gender and race influence self-ratings differentially: we expect women to rate themselves more harshly than men but we do not expect a race difference to emerge in self-ratings.

The quasi-exogenous shock where self-ratings were not shared with managers affords an identification strategy of potential social influence that theoretically should not be troubled by endogeneity concerns. We expect de-anchoring during the non-standard year to benefit women more than men, as we expect women to be more likely than men to underestimate their performance during the self-rating process (or vice versa, men more likely than women to overestimate their performance). The same does not apply for race: since self-ratings are not expected to vary by race, de-anchoring should not change the race dynamics.

Potential caveats to our identification strategy include specific year effects where the intervention year differs in ways other than the fact that self-evaluations were not shared. We compare the variables of interest between standard and non-standard years and do not detect any concerning evidence here. We also include a number of fixed effects that hold constant variables that might be correlated with differences in performance ratings, including geographic regions, job levels and idiosyncratic manager behaviors across years. Our results are generally robust across a wide range of empirical specifications, as we show below. In addition, the intervention

might not be strong enough to eliminate all social influence channels leading from employee self-evaluations to manager ratings. Specifically, managers might access employee performance scores from earlier years, which themselves might have been affected by previous years' employee self-evaluations. Our data suggests that this is a real concern, and we thus later focus on a specific sub-sample where this concern is—by construction—eliminated: employees with no history in the firm who receive their first performance rating in this firm during the non-standard year, the newcomers.

3. Field Context

3.1. Sample

We study the performance data of a global financial services company[5] over four years, from 2015-2018. The firm is headquartered in the United States and has offices in over 20 countries worldwide. Over this time period, about 60% of the employees identified as male, 40% as female, 43% as White, 32% as a person of color and 25% did not disclose their race[6] (Table AI).

3.2. Performance appraisal system

In this firm, performance is evaluated on a categorical scale with five choices, from "needing improvement" (codes as "1") to "significantly outperforming" (coded as "5").[7] In

[5] Due to the sensitive nature of this data, the firm context provided is limited as our data sharing agreement with the firm requires that the anonymity of the firm be protected.
[6] Disclosure rates of race varied substantially by region, varying from more than 97% disclosed in the Americas to 59% disclosed in EMEA (Europe-Middle East-Africa) to 45% disclosed in APAC (Asia-Pacific), reflecting variations in legal requirements or cultural norms related to identification by and the disclosure of race.
[7] The guidance provided to managers gives an overview of what is expected for each category. The highest rating is reserved for employees who significantly exceed high expectations and set new standards for the firm. The next highest rating is reserved for employees who have exceeded expectations and outperformed their peers, while the middle rating is for employees who meet or occasionally exceed expectations. The bottom two ratings are for employees who meet most but not all expectations, or fail to meet those expectations and are unlikely to improve.

assigning ratings to their employees, managers were encouraged to distribute the ratings such that only 5% of employees received the highest (5) and lowest (1) scores each. Managers were set targets of about 25% of employees receiving a score of 4, 45% receiving a score of 3, and 20% receiving a score of 2. Figure I shows that, in aggregate, managers generally produced a distribution of ratings that fit this pattern (blue bars). In contrast, on average, employees (red bars) were more optimistic about their own performance than their managers.

Insert Figure 1

In this firm, performance scores, in conjunction with managers' written assessments (which are not available to us), are collected once per year and are used to inform a number of talent decisions, including selection for leadership development programs, promotion readiness and year-end bonuses. The annual year-end performance management process typically begins in early October. Employees complete their self-assessment by end of October and managers assign performance ratings by end of November. Most managers meet with their team members in December go over the performance review and share the finalized, calibrated performance ratings.[8] The final deadline for completion of these conversations tends to be the first week in January. Employees receive one overall rating based on a five-point rating scale given by their direct manager, although the latter may have sought input from others, including direct reports, peers and other managers (on which we do not have information).

3.3. Standard and non-standard years

[8] Calibration meetings are a black box process to us. As we do not have any information about what happens in calibration meetings, we cannot rule out that they have an impact on the gender and race dynamics. That said, to the extent that this is the case, it is still notable that our results suggest that the calibration process is not designed (or able) to remove all gender or race differences. Furthermore, the evidence based on our results that de-anchoring lowers manager ratings further suggests that calibration meetings only play a limited role in reducing the impact of self-ratings.

In most years (i.e., the "standard" years), employees submitted their self-ratings before managers evaluated their employees, and followed the process as described above. We conduct our analyses for three of those years, 2015, 2017 and 2018, for which we have data and during which the performance evaluation process stayed the same. While all other aspects of the process remained the same, in 2016 (which we refer to as the "non-standard" year), the firm was unable to share employees' self-evaluations with managers beforehand. In our interviews with the firm, managers referred to a time crunch due to other factors, not allowing them to follow standard procedure, and assured us that this was not an intentional move. We examine performance scores by employees and managers before, during and after this quasi-shock over the four years where the other features of the process remained unchanged, from 2015-2018.

3.4. The global context and the US

We start by focusing on the global dataset for all countries in which the firm operates and employ binary definitions of gender and race comparing men and women as well as Whites and "people of color." To analyze race at a more granular level, including interactions between gender and different racial groups, we then take a closer look at the firm's home market, the United States, which comprises about half of the workforce.

Race data is incomplete in most countries in our dataset. The reasons for why race data is missing in some countries vary, including differences in legal requirements or, conversely, restrictions to collect race data from employees, as well as cultural norms where people are not used to defining themselves based on race or prefer not disclosing race, potentially introducing selection effects. The only country in our dataset with almost perfectly complete data on race is the United States. In addition to having an unbiased sample available in the US, we can also analyze various race categories separately, moving beyond a binary definition of race.

4. Results: All Countries

Based on the complete global data set, we discuss our results by first focusing on the baseline, standard years, where self-evaluations were shared with managers before they assessed the employees, and then turning to the intervention, the non-standard year, where they were not.

4.1. Standard years

We present summary statistics of self-ratings and manager ratings, split by employees' gender and race and their interaction for all countries across the standard years, 2015, 2017 and 2018, in Table A.2 in the Appendix. Figure 2 summarizes our findings for both self-ratings and manager ratings, which we test more rigorously below. We note three key takeaways from Figure 2: First, managers ratings are lower than self-ratings for all subgroups; second, race gaps appear in manager ratings while, third, gender gaps seem to be most pronounced for self-ratings. We unpack these findings in the following sections.

Insert Figure 2

4.1.1. Manager ratings. Table 1 examines manager ratings by demographic group in standard years more precisely. In order to introduce the notion and regression structure for the remainder of the paper, Columns 1 and 2 show the estimates for the following simple ordinary least squares (OLS) models, respectively:

$$m_{ijlgt} = \beta_0 + \beta_1 f_i + \varepsilon_j \quad (1)$$

$$m_{ijlgt} = \beta_0 + \beta_1 r_i + \varepsilon_j \quad (2)$$

The dependent variable, m_{ijlgt}, is an ordinal performance rating of employee *i* by manager *j* in job level *l* in geographic region *g* in year *t*. The manager rating ranges from 1 to 5, where a higher number indicates the manager's impression of a better performance of the employee. The dummy variables, f_i and r_i, indicate the gender and race of employee *i*, respectively: f_i is 1 if the employee is female and 0 otherwise, and r_i is 1 if the employee is a person of color (i.e. any self-selected category other than "White") and 0 otherwise. There are five job-levels (l): administrative assistant, junior level, middle management, junior senior management and senior management. There are three geographic regions (g): the Americas, APAC (Asia-Pacific), and EMEA (Europe, Middle East, Africa). Robust standard errors (ε_j) are clustered at the manager level.

When only examining gender or race, respectively, we see that female employees received significantly lower ratings than male employees (Table 1, Column 1) and employees of color received significantly lower ratings than White employees (Column 2). However, once we account for gender and race together, the gender gap disappears while the magnitude of the race gap remains relatively unchanged (Column 3). In Column 4, we include interaction effects (i.e., being a woman of color): while the gender gap remains not significant for White women, a significant interaction emerges: the race gap is even larger for women than for men.

To ensure robustness of these findings, we add several fixed effects. Manager (θ_j), geographic region (θ_g), job level (θ_l) and year (θ_t) fixed effects are included, as rating standards may differ by individual managers, across geographic regions, between job levels and across years. Equation (3) presents the full model:

$$m_{ijlgt} = \beta_0 + \beta_1 f_i + \beta_2 r_i + \beta_3 f_i r_i + \theta_j + \theta_g + \theta_l + \theta_t + \varepsilon_j \quad (3)$$

Table 1 presents several fixed-effects estimates in Columns 5-8. Once we control for iteratively more variables, including manager fixed effects (Column 5), regional fixed effects (Column 6), job-level fixed effects (Column 7) and year fixed effects (Column 8), the gender gap among White employees reverses with White women being evaluated more favorably than White men. The gender-race interaction remains significant across specifications: the race gap is larger for female than for male employees. Across the fixed-effects models, all effect sizes remain relatively similar.

Insert Table 1

Table A.3 in the Appendix replicates our analysis interacting employee demographics with manager demographics, finding consistent race effects independent of the manager's race and some nuances in the gender dynamics with female managers assigning lower ratings than male managers, in particular to male employees.

4.1.2. Self-ratings. Manager ratings may be due to managers' own independent assessment and/or influenced by employees' self-ratings. We take a look at employees' self-evaluations and their relationship to gender and race in Table 2 repeating the same analyses as above with self-ratings on the left-hand side of the regression.

Female employees gave themselves significantly lower ratings than male employees (Column 1) as did employees of color compared with their White counterparts (Column 2). When we account for gender and race simultaneously, we see the same pattern (Column 3); however, once we control for the interaction between gender and race, we find that the apparent race gap in self-ratings was driven entirely by women of color (Column 4). There is no significant difference in self-ratings between White men and men of color. These findings

remain relatively constant, once managerial, regional, job-level and year fixed effects are accounted for (Columns 5-8).

Insert Table 2

4.1.3. Manager ratings and self-ratings. In Table A.4 in the Appendix, we show that manager ratings are consistently correlated with self-ratings, although they are on average lower than self-ratings. This association may be the result of either independent, directionally similar assessments by the manager and employee, or a causal relationship where the employee's self-rating anchors the manager, or both.

Even when controlling for self-ratings in Table A.4, gender and race effects remain important: as can be seen in Figure 2, the difference between self- and manager ratings is smaller for female than for male employees (suggesting managers reversed the gender gap in self-ratings) and larger for employees of color than for White employees (suggesting managers introduced a race gap that was not present for self-ratings).

However, when using this regression framework in standard years, we cannot distinguish whether managers were influenced by the gender and race of the employee directly, or through the self-ratings which, in turn, were influenced by these demographic characteristics (see Table 2). We therefore turn to a quasi-exogenous shock to the supply of self-ratings to understand what, if any, the role of social influence was that employees provided to managers via self-ratings.

4.2. The non-standard year

4.2.1. Quasi-exogenous shock. In 2016, the firm experienced a quasi-exogenous shock, whereby managers had to provide their ratings before having the opportunity to view employees' self-ratings. While not an experiment, balancing tests of our samples in the two time periods of interest (Table A.5 in the Appendix) do not suggest dramatic differences in the composition of employees in standard years as compared to the non-standard year. In addition, self-ratings in the standard years and the non-standard year did not differ on average, suggesting that employees were indeed unaware of the change and also did not change their own behavior in other unexpected ways (even if doing so would not actually have been observed by their manager before they made up their minds). Nonetheless, as it was not a randomized controlled trial, general time trends or unobserved characteristics might still affect our results.

4.2.2. De-anchoring. Figure 3 shows that, in general, manager ratings were lower in the non-standard year than in standard years, which suggests that de-anchoring occurred. At the same time, however, gender and race dynamics seem to be the same across all years.

Insert Figure 3

In Table 3, we examine de-anchoring econometrically. Column 1 confirms that managers assigned lower ratings to employees in the non-standard than in standard years. We interpret this significant, negative coefficient as an indication that the presence of self-ratings in standard years typically influenced managers' decision-making regarding the final rating to some extent. Put differently, when the employee-supplied rating was removed (i.e. self-ratings were not shown in the non-standard year), managers assigned lower ratings to all employees.

4.2.3. Gender and race dynamics. Turning to demographic variables in Table 3, the non-standard year showed similar gender and race effects as in the standard years, independent of the controls we add (Columns 2-6), suggesting that the kinds of behaviors managers exhibited in standard years also carried over to the non-standard year. Even when managers did not have employee self-evaluations available in 2016, they assigned higher ratings to White female employees than White male employees, and lower ratings to all employees of color, with the race gap being most pronounced for women.

Insert Table 3

4.2.4. Shadow of the past. Given that managers' ratings were significantly lower in the non-standard year, we conjecture that self-ratings might have had some influence on managers' ratings in standard years. To explore the dynamics in the non-standard year further, we consider that most managers and employees have a history (including of self-ratings in previous years potentially having influenced manager ratings in previous years). While our data does not allow us to make causal inferences due to this inherent endogeneity over time, we expect that such a history might matter, especially in the non-standard year. Because managers do not have access to self-ratings in the non-standard year, managers could feel inclined to rely on other information (i.e., another anchor), including their own rating of the employee the year before – the "shadow of the past." While we do not have data to know whether a manager in fact consulted prior ratings, we would expect that, if they had, their manager rating would be more correlated with their own manager rating from the previous year in the non-standard than in the standard years

(and, because self-ratings are not observable, less correlated with employee's current self-ratings in the non-standard year).

Indeed, the raw correlation coefficients between manager and self-ratings are $r_s = 0.46$ in standard years and $r_{ns} = 0.41$ in the non-standard year. In contrast, manager ratings appear less correlated with their previous year's own rating in standard years ($r_s = 0.38$) than in the non-standard year ($r_{ns} = 0.43$). Using our standard regression framework, Table A.6 in the Appendix shows that these differences are significant. Managers indeed seem to have relied more on past ratings in the year where self-evaluations were not available than in standard years.

4.2.5. Newcomers. Because the "shadow of the past" makes current year's ratings endogenous, we now focus on a group of employees where the shadow of the past can be ruled out by construction. Specifically, we take a closer look at employees' ratings during their first year of employment in the company, the newcomers, conducting subgroup analyses for the standard year and the non-standard year (Table A.7 in the Appendix). First, we note again that manager ratings in the non-standard year were lower than in the standard year, suggesting evidence of de-anchoring (as before for all employees). Second, while we find a persistent race gap for all newcomers in standard years (as was the case for all employees), the race gap disappeared for women in the non-standard year, and women of color were rated on par with both male and female White employees (*unlike* for all employees). In contrast to standard years, the additional disadvantage women of color experienced due to their lower self-evaluations was removed by the de-anchoring intervention in the non-standard year, reversing the gender gap among newcomers of color such that, in this case, male newcomers of color ended up receiving the lowest ratings.

However, while directionally consistent, we do not have sufficient power to statistically document this effect for women of color among newcomers in a fully-specified triple interaction (the coefficient on female x people of color x non-standard year is positive but n.s. with p=0.14). We thus take the newcomer results only as suggestive evidence that managers might assign higher ratings to female newcomers of color when not anchored by their self-evaluations. Figure 4 illustrates the effects for newcomers graphically.

Insert Figure 4

5. Results: United States

Our analysis so far has included the entire dataset across all countries and years. Our global results suggest persistent gender effects for self-evaluations and race effects for manager evaluations.

A closer look at our data reveals that the race gap is heavily driven by the US. In Table A.8 in the Appendix, we show how our main findings break down when looking at the US and other countries separately. While many dynamics seem to apply across geographies—e.g., managers reversing the gender gap in self-ratings—the manager driven race dynamics appear much more pronounced in the US. We now dissect the US effects further, and also run counter-factual simulations to better understand the resulting effect sizes.

5.1. Standard Years and Non-Standard Year in the US

We follow the same econometric strategy (see Equations 1-3) to first analyze manager ratings (see Table A.11 in the Appendix), then self-ratings (Table A.12) and finally their

relationship (Table A.13) in the US dataset. Most of our results in the US mirror our global findings, with only a few exceptions and nuances which we summarize here.[9]

5.1.1. Standard Years. For manager ratings, while all employees of color received lower ratings than their White counterparts, the effect was most pronounced for Black employees who received the most negative ratings independent of our specification. The gender dynamics in the US were directionally similar to the pattern for the world but no longer significant. For self-ratings, the gender gap seemed to be mostly driven by Asian Americans, with Asian American women giving themselves lower ratings than their male counterparts. In addition, a race gap not observed in the world data emerged with Black employees giving themselves lower ratings than their White counterparts.[10] When controlling for self-ratings in standard years, we again find that gender and race effects remained important. As with the global dataset, managers in the US reversed the gender gap but introduced a race gap. Black employees received the lowest manager ratings.

5.1.2. Non-standard year. While manager ratings were lower in the non-standard year than in the standard year, suggesting "de-anchoring" took place, the race and gender dynamics remained unaffected by the intervention. As before, managers in the US already addressed the gender gap in self-ratings in standard years, which meant that the non-standard year did not lead to any substantive changes. Meanwhile the race gap in manager ratings was unaffected by the intervention, as it mostly was induced by managers. The only exception were Black employees whose self-ratings were lower compared to White employees. However, as in the global sample,

[9] For summary statistics of the average self-ratings and manager ratings for the five racial groups we can distinguish in the US (Asian, Black, Latinx, Other and White Americans), see Table A.9 in the Appendix. Table A.10 shows average ratings for all possible gender and race combinations.

[10] This is the first and only time we observe a race gap in self-ratings, which seems to be unique to the US context for Black employees. To the best of our knowledge, a race gap in self-ratings has not previously been reported.

the intervention did not differentially benefit those who gave themselves lower self-ratings, likely because managers in the US also had prior year ratings available.[11]

5.2. Race Simulations for the US

Our results suggest that the race gaps in the US were primarily driven by managers. To better illustrate the magnitude of the effect sizes we observe for race in the US, we conduct a number of counterfactual simulations. We run bootstrapped simulations, drawing from the original data, with the goal of identifying how many Asian, Black or Latinx employees would have to receive a more positive manager rating for us to no longer observe differences between demographic groups. These simulations enable us to study the magnitude of the observed effects. We only draw from observations in standard years where self-ratings were observable to managers. We estimate a model that includes dummy variables for gender, all racial categories and a number of fixed effects:

$$m_{ijlgt} = \beta_0 + \beta_1 f_i + \beta_2 r_{A,i} + \beta_3 r_{B,i} + \beta_4 r_{L,i} + \beta_4 r_{O,i} + \theta_j + \theta_g + \theta_l + \theta_t + \varepsilon_j \qquad (4)$$

where $r_{A,i}$ is 1 if the employee is Asian; $r_{B,i}$ is 1 if the employee is Black; $r_{L,i}$ is 1 if the employee is Latinx; $r_{O,i}$ is 1 if the employee is in the "Other" category; all other variables are as defined above for Eqs. (1-3).[12] As such, we explore how many employees of each racial category—regardless of their gender—would need to experience a higher rating (i.e., an increase of 1 unit on a scale from 1 to 5), so that the average rating of that group is indistinguishable from White employees.

[11] While an analysis similar to our global approach focusing on newcomers would also be interesting in the US context, we unfortunately do not have enough sample size to warrant such an investigation: there were only 45 Black newcomers in the US in the non-standard year (out of a total of 944).

[12] We chose not to include the interaction terms between gender and racial categories because the policy-relevant counterfactual does not require intersectionality in the United States data: as Table A.10 shows, men and women of color (including Asian, Black, Latinx and "Other" employees) are experiencing lower manager ratings than White employees regardless of their gender.

In the first set of simulations, we increase the manager rating of a randomly selected subset of Black employees. We focus on Black employees because the race gap in manager ratings is most pronounced for them. We conduct 100 iterations for each fraction of the subset: in each iteration, we draw a subset of Black employees without replacement, increasing the manager rating by 1 unit (unless the employees had already received the highest manager rating[13]), estimate Eq. (4) and save the regression coefficients and standard errors associated with $r_{B,i}$. After all iterations, we calculate the mean coefficient $\hat{\beta}_3$, the mean standard error, as well as the associated t-statistic and *p*-value. This process is repeated for differently sized fractions of Black employees to identify above which threshold there exists no difference in manager ratings between White employees and Black employees.

Insert Table 4

Table 4 shows that approximately 22% of Black employees in the United States would need to experience a higher manager rating than they currently receive, in order for there to be no significant difference between the managers' ratings of White employees and Black employees ($p > 0.05$). Furthermore, beyond non-significance, we also explore at what point the coefficient is closely estimated at 0 (i.e. virtually no difference between the two groups), which would require a subset of at least 28% of Black employees to be affected by higher manager ratings. In additional simulations, we repeat the same process with Asian, Latinx and employees who self-selected into the "Other" racial category. In the Appendix, Tables A.15–A.17 show that 4-7%

[13] For a small fraction of employees of color (who have the highest manager rating prior to the simulation change) the manager rating will not be altered, although they are technically "treated". Excluding employees with a rating of 5 from this procedure does not affect our results or the conclusions we can draw (results not shown).

Asian, 4-11% Latinx and 3-12% "Other" employees would need to experience an increase in their manager rating for there to be no difference to White employees.

6. Discussion and Conclusion

Performance reviews are prone to allegations of bias, and some fear that remote work, prevalent for many during the Covid-19 pandemic, might have exacerbated bias in such reviews (Mackenzie et al. 2019; Lanik 2020). News reports suggest that some companies have taken action. For example, Amazon announced in April 2021 that it would "inspect any statistically significant demographic differences in Q1 2021 performance ratings … to identify root causes and, as necessary, implement action plans" (Galetti 2021). Clearly, different interventions are called for depending on the causes of the observed differences.

Working with a financial services firm headquartered in the United States, we evaluate the impact of one particular intervention: not sharing employees' self-ratings with managers before the latter evaluated their employees. We focus on the two demographic characteristics available to us, gender and race, and their interaction. If differences in final performance ratings were primarily due to differences in self-evaluations, not sharing self-evaluations could interrupt this social influence channel leading to anchoring. Alternatively, if differences in final scores were mostly driven by managers, independent of employees' self-ratings, this intervention might have little impact on observed gender or race dynamics.

We find evidence for de-anchoring—but it had only limited impact on the gender and race patterns in this firm: managers assigned lower performance ratings when employees' self-ratings were not available to them. This did not affect general gender or race dynamics for two reasons. While White women and even more so, women of color, assigned themselves lower

self-ratings than their respective male counterparts, managers closed these gaps for White women even when self-ratings were shared. They did not do so for women of color but instead, added a race gap not present in self-ratings, leaving female employees of color with the worst final ratings overall. In addition, managers had previous years' ratings available even when self-ratings were not shared, adding possible historical anchors. To remove the latter, we focus on employees with no history in the firm: newcomers during their first year of employment. Among these newcomers, employees who had traditionally assigned themselves the lowest self-ratings, women of color, benefited from the removal of their self-evaluations, ending up with final ratings on par with White employees and higher than their male counterparts, thus reversing the gender gap typically introduced by self-ratings also for people of color.

While our sample size for newcomers is too small to derive firm conclusions, we take our results as suggestive evidence supporting the notion that not sharing self-evaluations with managers before they make up their minds in performance appraisal systems could help level the playing field for those most negatively affected by their self-evaluations – assuming they do not already have an established history with the manager. In this firm, these most negatively affected employees were female employees of color. This group might have experienced "double jeopardy" where both gender and race effects led women of color to assign themselves lower self-evaluations.

Our results also suggest that we should not expect dynamics not related to self-evaluations to be impacted by this intervention. In this firm, managers introduced most of the race gap present in final ratings, in particular in the US where about a quarter of Black employees would have to receive better scores for the race gap to be closed. The race gaps were also significant but smaller for Latinx and Asian employees in the US. Race gaps remained

relatively stable across the years, independent of whether self-evaluations were shared or not (with the exception of the newcomer effects discussed above).

Subjective performance appraisals similar to the ones analyzed in this firm are common in most firms as objective performance data is rarely available for complex jobs. While neither the firm nor we can assess to what degree differences in final performance ratings are due to true underlying differences in performance or to self- or manager bias, systematic differences based on demographic characteristics are of concern as these performance scores are typically used to inform compensation and career advancement decisions and thus could induce systemic inequities. Firms like Amazon, thus, are well advised to take a closer look at their performance ratings "to identify root causes," as they write, for potential differences by demographic characteristics.

Our paper contributes to the on-going debate about how such inequities could be addressed, suggesting that differentiating between employees' self-evaluations, managers' evaluations and their interplay could be useful in informing policy. If employees' self-evaluations were heavily affected by demographic characteristics, e.g., through self-stereotyping or expectations of social backlash, and then anchored the managers, interventions aimed at employees or at disabling anchoring would be particularly fruitful. If alternatively, managers were mostly responsible for the gender and race differences in final performance scores, then interventions focused on managers would be called for.

References

Arnold, D., Dobbie, W.S. and Hull, P., 2020. Measuring racial discrimination in bail decisions (No. w26999). *National Bureau of Economic Research.*

Arrow, K.J., 1973. *Information and Economic Behavior*. Cambridge: Harvard University Press.

Babcock, L. and Laschever, S., 2003. *Women Don't Ask: Negotiation and the Gender Divide*. Princeton University Press.

Barber, B.M. and Odean, T., 2001. Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1), pp.261-292.

Beale, F.M., 1970. *Black women's manifesto; double jeopardy: To be Black and female*. New York: Third World Women's Alliance.

Berdahl, J.L. and Moore, C., 2006. Workplace harassment: double jeopardy for minority women. *Journal of Applied Psychology*, 91(2), p.426.

Bertrand, M. and Duflo, E., 2017. Field experiments on discrimination. In: *Handbook of economic field experiments* (Vol. 1, pp. 309-393). North-Holland.

Bertrand, M. and Mullainathan, S., 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4), pp.991-1013.

Biernat, M. and Sesko, A.K., 2013. Evaluating the contributions of members of mixed-sex work teams: Race and gender matter. *Journal of Experimental Social Psychology*, *49*(3), pp.471-476.

Bohnet, I., 2016. *What Works: Gender Equality by Design*. Cambridge: Harvard University Press.

Bohnet, I., Van Geen, A. and Bazerman, M., 2016. When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, 62(5), pp.1225-1234.

Bohren, J. A., Haggag, K., Imas, A. and Pope, D. G., 2019. Inaccurate statistical discrimination (No. w25935). *National Bureau of Economic Research*.

Bohren, J.A., Imas, A. and Rosenberg, M., 2019. The dynamics of discrimination: Theory and evidence. *American Economic Review*, 109(10), pp.3395-3436.

Bordalo, P., Coffman, K., Gennaioli, N. and Shleifer, A., 2016. Stereotypes. *The Quarterly Journal of Economics*, 131(4), 1753-1794.

Bordalo, P., Coffman, K., Gennaioli, N. and Shleifer, A., 2019. Beliefs about gender. *American Economic Review*, 109(3), 739-73.

Bosquet, C., Combes, P.P. and García-Peñalosa, C., 2019. Gender and promotions: evidence from academic economists in France. *The Scandinavian Journal of Economics*, 121(3), pp.1020-1053.

Bowles, H.R., Babcock, L. and Lai, L., 2007. Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask. *Organizational Behavior and Human Decision Processes*, 103(1), pp.84-103.

Buser, T., Niederle, M. and Oosterbeek, H., 2014. Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, 129(3), 1409-1447.

Castilla, E.J., 2008. Gender, race, and meritocracy in organizational careers. *American Journal of Sociology*, 113(6), pp.1479-1526.

Castilla, E.J., 2012. Gender, race, and the new (merit-based) employment relationship. *Industrial Relations: A Journal of Economy and Society*, *51*, pp.528-562.

Castilla, E.J., 2015. Accounting for the gap: A firm study manipulating organizational

accountability and transparency in pay decisions. *Organization Science*, 26(2), pp.311-333.

Coffman, K.B., 2014. Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4), pp.1625-1660.

Coffman, K. B., Exley, C. L. and Niederle, M., 2021. The Role of Beliefs in Driving Gender Discrimination. *Management Science*.

Crenshaw, K.W., 2017. *On intersectionality: Essential writings*. The New Press.

Croson, R. and Gneezy, U., 2009. Gender differences in preferences. *Journal of Economic Literature*, 47(2), pp.448-74.

DeNisi, A. S. and Murphy, K. R., 2017. Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology*, 102(3), pp.421–433.

Dobbin, F., Schrage, D. and Kalev, A., 2015. Rage against the iron cage: The varied effects of bureaucratic personnel reforms on diversity. *American Sociological Review*, 80(5), pp.1014-1044.

Dulebohn, J. H. (1995). *Social influence and organizational justice in employee reactions to performance appraisals* (Doctoral dissertation, University of Illinois at Urbana-Champaign).

Dulebohn, J. H., and Ferris, G. R. (1999). The role of influence tactics in perceptions of performance evaluations' fairness. *Academy of Management journal*, *42*(3), 288-303.

Exley, C.L. and Kessler, J.B., 2019. The gender gap in self-promotion (No. w26345). *National Bureau of Economic Research*.

Furnham, A. and Boo, H.C., 2011. A literature review of the anchoring effect. *The Journal of Socio-economics*, 40(1), pp.35-42.

Galetti, B., 2021. *Diversity, Equity, and Inclusion*. Accessed on May 20, 2021 at https://www.aboutamazon.com/news/workplace/diversity-equity-and-inclusion

Gallus, J. and Heikensten, E., 2019. *Shine a light on the bright: The effect of awards on confidence to speak up in gender-typed knowledge work*. Working Paper.

Gallus, J. and Heikensten, E., 2020, May. Awards and the Gender Gap in Knowledge Contributions in STEM. In *AEA Papers and Proceedings* (Vol. 110, pp. 241-44).

Glover, D., Pallais, A. and Pariente, W., 2017. Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. *The Quarterly Journal of Economics*, 132(3), pp.1219-1260.

Goldin, C. and Rouse, C., 2000. Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American Economic Review*, 90(4), pp.715-741.

Greenhaus, J. H., Parasuraman, S., and Wormley, W. M. (1990). Effects of race on organizational experiences, job performance evaluations, and career outcomes. *Academy of management Journal*, 33(1), 64-86

Hauser, D.N. and Bohren, J.A., 2021. Learning with Heterogeneous Misspecified Models: Characterization and Robustness.

Hospido, L., Laeven, L. and Lamo, A., 2019. The gender promotion gap: evidence from central banking. *Review of Economics and Statistics*.

Joshi, A., Son, J. and Roh, H., 2015. When can women close the gap? A meta-analytic test of sex differences in performance and rewards. *Academy of Management Journal,* 58(5), pp.1516-1545.

Klimoski, R. and Inks, L., 1990. Accountability forces in performance appraisal. *Organizational Behavior and Human Decision Processes*, 45(2), pp.194-208.

Korsgaard, M. A. (1996). The impact of self-appraisals on reactions to feedback from others: the role of self-enhancement and self-consistency concerns. *Journal of Organizational Behavior*, 17(4), 301-311.

Lanik, M., 2020. *Why This Year's Talent Reviews Are the Perfect Storm for Bias and Discrimination*. Accessed on May 20, 2021 at https://www.pinsight.com/blog/why-this-years-talent-reviews-are-the-perfect-storm-for-bias-and-discrimination/

Latham, G. P., Budworth, M. H., Yanar, B., and Whyte, G. (2008). The influence of a manager's own performance appraisal on the evaluation of others. *International Journal of Selection and Assessment*, *16*(3), 220-228.

Livingston, R.W., Rosette, A.S. and Washington, E.F., 2012. Can an agentic Black woman get ahead? The impact of race and interpersonal dominance on perceptions of female leaders. *Psychological science*, *23*(4), pp.354-358.

Mackenzie, L. N., Wehner, J. and Correll, S. J., 2019. Why Most Performance Evaluations Are Biased, and How to Fix Them. *Harvard Business Review*.

McKay, P.F. and McDaniel, M.A., 2006. A reexamination of black-white mean differences in work performance: More data, more moderators. *Journal of Applied Psychology*, 91(3).

Milkman, K.L., Akinola, M. and Chugh, D., 2015. What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology*, *100*(6), p.1678.

Moss-Racusin, C.A., Dovidio, J.F., Brescoll, V.L., Graham, M.J. and Handelsman, J., 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41), pp.16474-16479.

Niederle, M. and Vesterlund, L., 2007. Do women shy away from competition? Do men compete

too much? *The Quarterly Journal of Economics*, 122(3), pp.1067-1101.

Pager, D. and Pedulla, D.S., 2015. Race, self-selection, and the job search process. *American Journal of Sociology*, 120(4), pp.1005-1054.

Parsons, C.A., Sulaeman, J., Yates, M.C. and Hamermesh, D.S., 2011. Strike three: Discrimination, incentives, and evaluation. *American Economic Review*, 101(4), pp.1410-35.

Paustian-Underdahl, S. C., Walker, L. S., and Woehr, D. J. (2014). Gender and perceptions of leadership effectiveness: A meta-analysis of contextual moderators. *Journal of applied psychology*, *99*(6), 1129.

Phelan, J.E. and Rudman, L.A., 2010. Reactions to ethnic deviance: The role of backlash in racial stereotype maintenance. *Journal of Personality and Social Psychology*, 99(2).

Phelps, E.S., 1972. The statistical theory of discrimination. *American Economic Review*, 62(4), pp.659-661.

Price, J. and Wolfers, J., 2010. Racial discrimination among NBA referees. *The Quarterly Journal of Economics*, 125(4), pp.1859-1887.

Purdie-Vaughns, V. and Eibach, R.P., 2008. Intersectional invisibility: The distinctive advantages and disadvantages of multiple subordinate-group identities. *Sex roles*, *59*(5), pp.377-391.

Quadlin, N., 2018. The mark of a woman's record: Gender and academic performance in hiring. *American Sociological Review*, 83(2), pp.331-360.

Quillian, L., Pager, D., Hexel, O. and Midtbøen, A.H., 2017. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences*, 114(41), pp.10870-10875.

Ridgeway, C.L. and Kricheli-Katz, T., 2013. Intersecting cultural beliefs in social relations: Gender, race, and class binds and freedoms. *Gender & Society,* 27(3), pp.294-318.

Rosette, A.S. and Livingston, R.W., 2012. Failure is not an option for Black women: Effects of organizational performance on leaders with single versus dual-subordinate identities. *Journal of Experimental Social Psychology*, 48(5), pp.1162-1167.

Rosette, A.S., Akinola, M. and Ma, A., 2018. Subtle discrimination in the workplace: Individual-level factors and processes.

Roth, P.L., Huffcutt, A.I. and Bobko, P., 2003. Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology*, 88(4).

Schleicher, D. J., Baumann, H. M., Sullivan, D. W., and Yim, J. (2019). Evaluating the effectiveness of performance management: A 30-year integrative conceptual review. Journal of Applied Psychology, 104(7), 851.

Shore, T.H., Adams, J.S. and Tashchian, A., 1998. Effects of self-appraisal information, appraisal purpose, and feedback target on performance appraisal ratings. *Journal of Business and Psychology*, 12(3), pp.283-298.

Society for Human Resource Management, 2014. *HR Professionals' Perceptions About Performance Management Effectiveness*. Accessed on 7 July 2019 at https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/pages/2014-performance-management.aspx.

Thorsteinson, T.J., Breier, J., Atwell, A., Hamilton, C. and Privette, M., 2008. Anchoring effects on performance judgments. *Organizational Behavior and Human Decision Processes*, 107(1), pp.29-40.

Tversky, A. and Kahneman, D., 1974. Judgment under uncertainty: Heuristics and biases.

Science, 185(4157), pp.1124-1131.

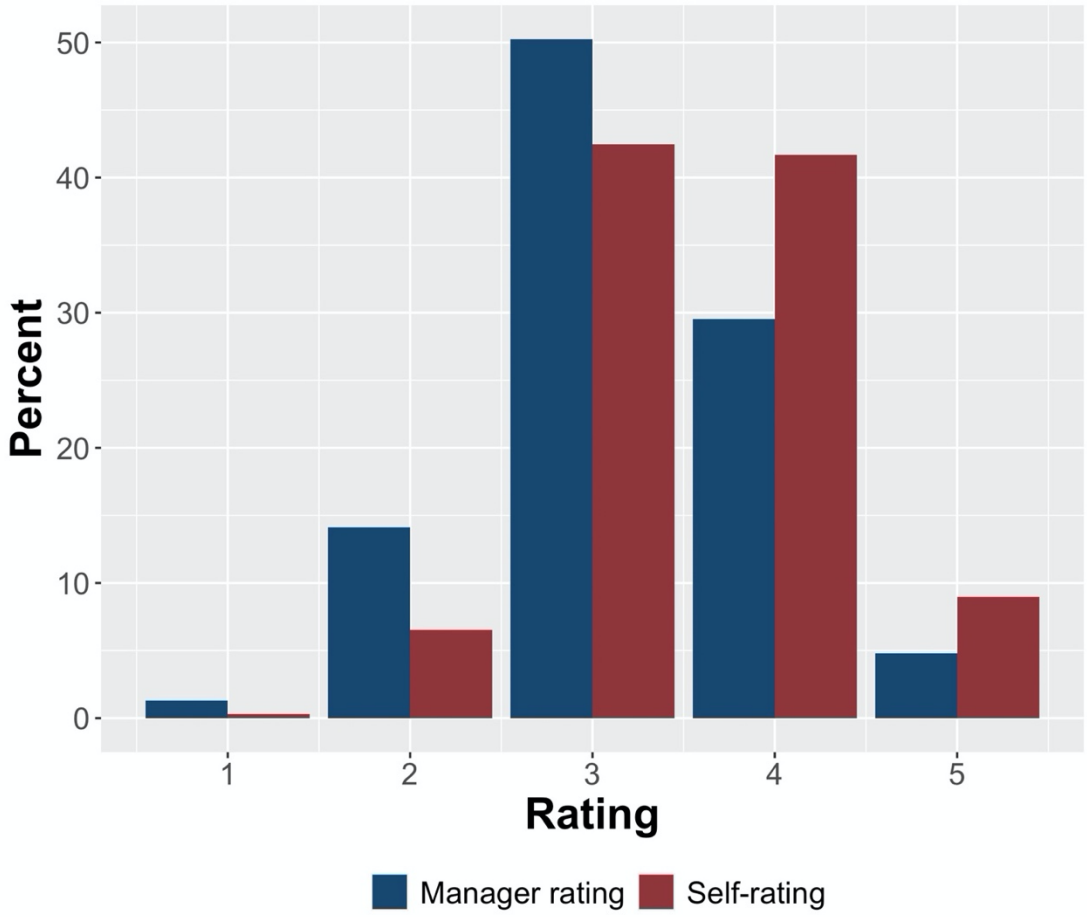

Figure 1.
Distribution of Ratings in Standard Years

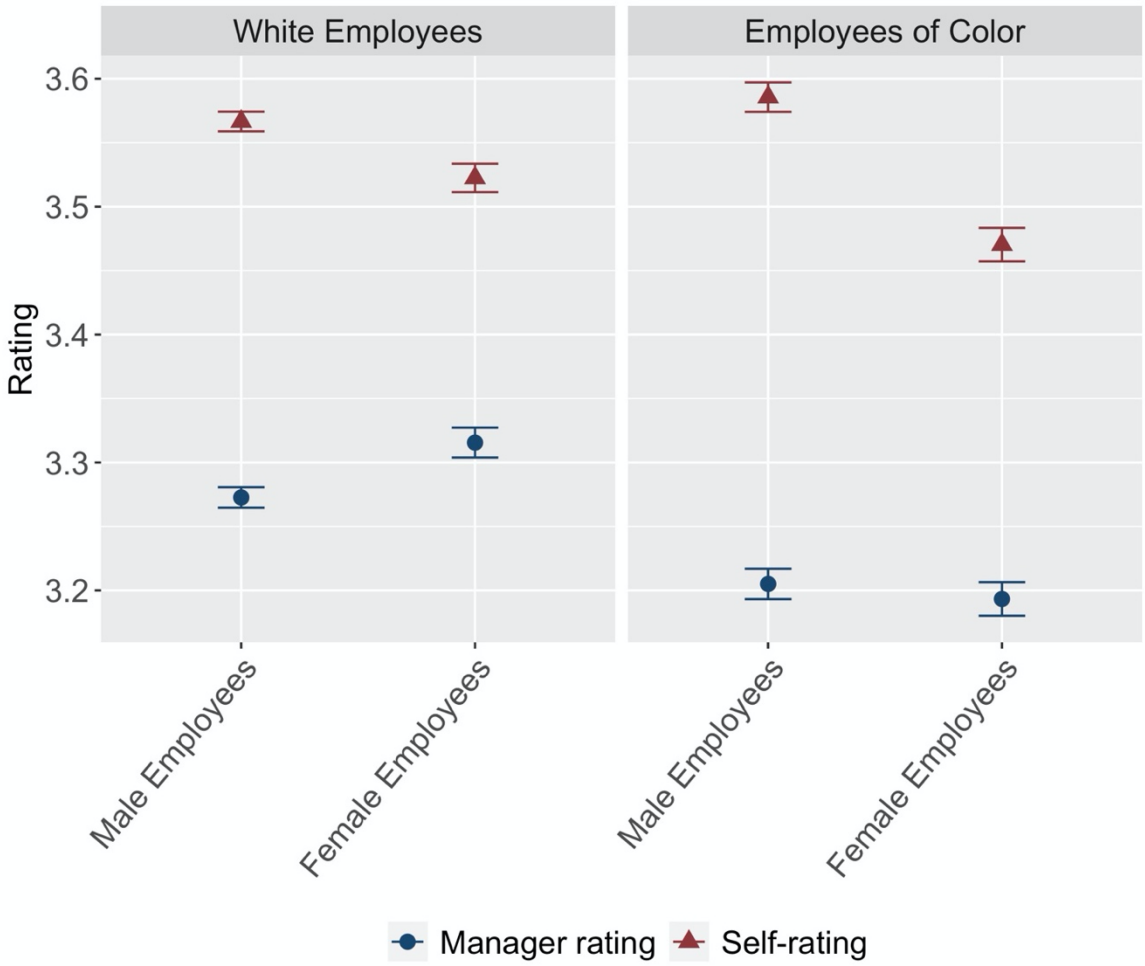

Figure 2. Estimated self-ratings and manager ratings by race and gender

Notes. This figure shows manager ratings (blue dots) and self-ratings (red triangles) across all countries in standard years, controlling for all fixed-effects in Column 8 in Table 1 (manager ratings) and Table 2 (self-ratings), respectively. Compared to self-ratings, manager ratings across all subgroups are lower. The difference between self- and manager ratings is smaller for female than for male employees, and larger for employees of color than for White employees. This reverses the gender gap present in self-ratings in final ratings for White employees with White women receiving higher ratings than White men and closes it for employees of color, but it also introduces a race gap with all employees of color ending up with lower scores than their White counterparts. Error bars represent standard errors from the mean.

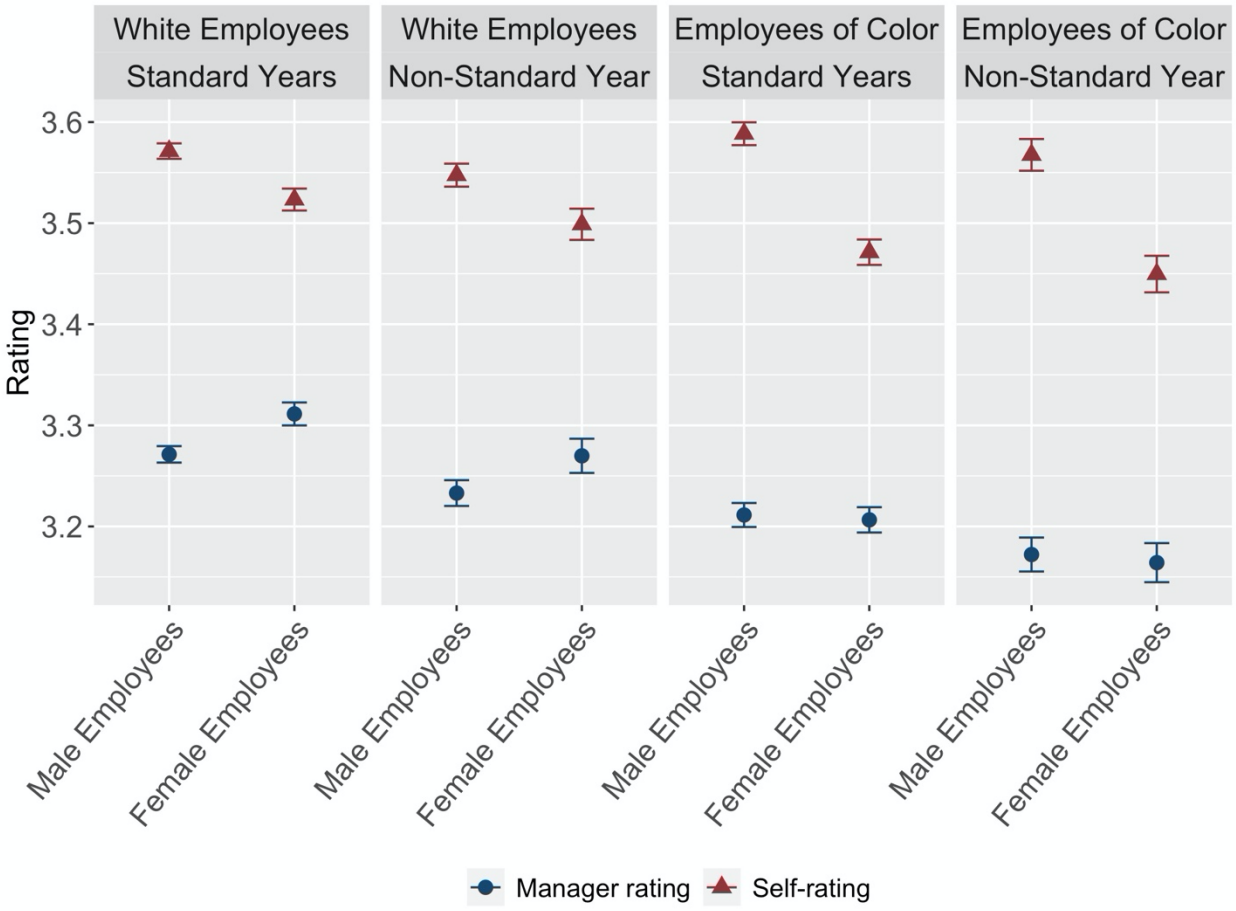

Figure 3. Estimated manager and self-ratings in standard and non-standard year, by gender and race

Notes. This figure shows manager ratings (blue dots) and self-ratings (red triangles) across all countries in standard years versus the non-standard year, controlling for all fixed-effects. Compared to standard years, manager ratings across all subgroups are lower in the non-standard years, suggesting "de-anchoring." However, the gender and race dynamics are largely unchanged between the standard years and the non-standard year. Error bars represent standard errors from the mean.

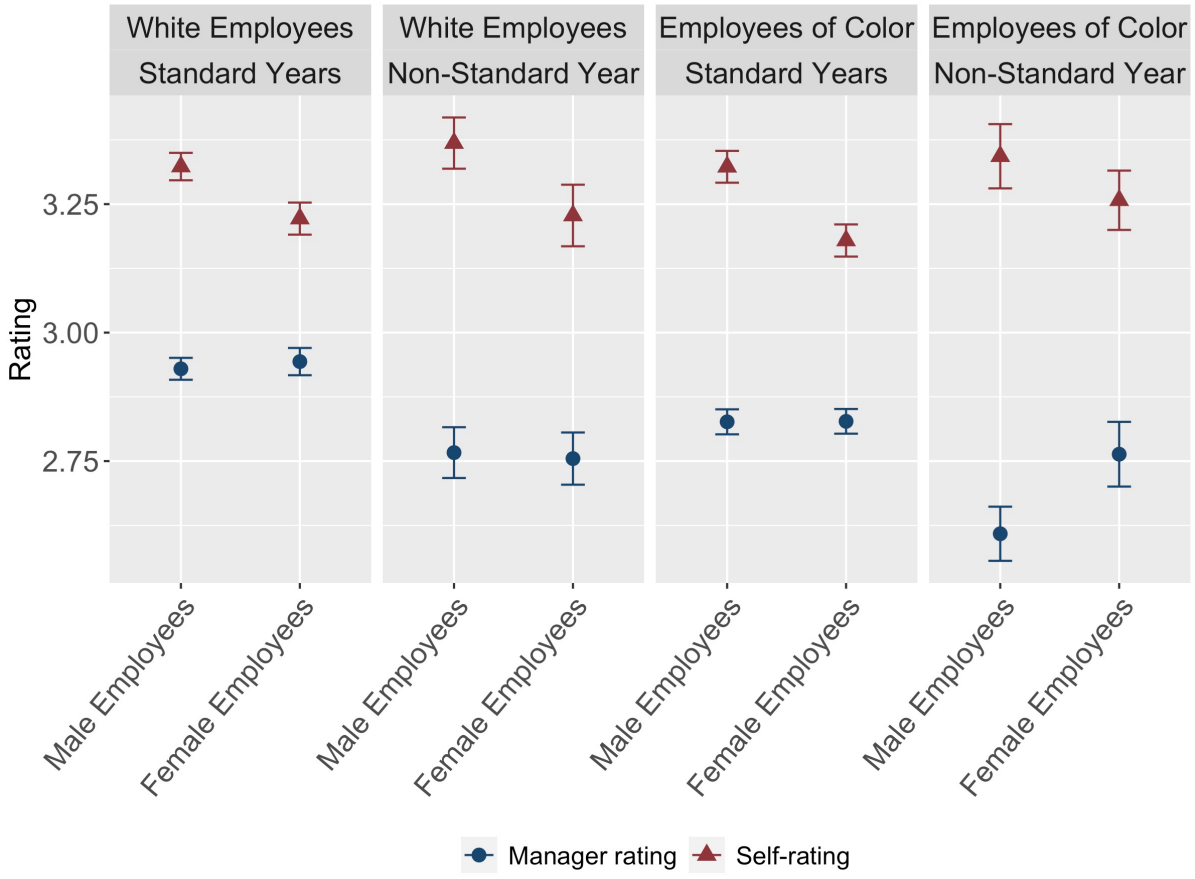

Figure 4. Newcomers: Estimated manager and self-ratings in standard and non-standard year, by gender and race

Notes. This figure shows manager ratings (blue dots) and self-ratings (red triangles) for newcomers only, across all countries in standard years versus the non-standard year, controlling for all fixed-effects. Unlike in the case of all employees, the de-anchoring intervention in the non-standard year led to a notable difference for one subgroup: women of color received higher manager ratings in the non-standard year, ending up with scores similar to White men and women, whereas men of color received the lowest manager ratings. Error bars represent standard errors from the mean.

Table 1. Manager Ratings in All Countries in Standard Years

	(1) Manager Rating	(2) Manager Rating	(3) Manager Rating	(4) Manager Rating	(5) Manager Rating	(6) Manager Rating	(7) Manager Rating	(8) Manager Rating
Female	-0.031**		-0.009	0.022	0.043**	0.043**	0.048**	0.043**
	(0.009)		(0.011)	(0.014)	(0.016)	(0.016)	(0.016)	(0.016)
People of Color		-0.082***	-0.082***	-0.053***	-0.052**	-0.057***	-0.063***	-0.068***
		(0.011)	(0.011)	(0.014)	(0.017)	(0.017)	(0.017)	(0.017)
Female*People				-0.073***	-0.063**	-0.060*	-0.056*	-0.055*
of Color				(0.021)	(0.024)	(0.024)	(0.024)	(0.024)
Constant	3.235***	3.282***	3.286***	3.275***	3.268***	3.270***	3.269***	3.273***
	(0.007)	(0.008)	(0.009)	(0.010)	(0.008)	(0.008)	(0.008)	(0.008)
Manager FE	N	N	N	N	Y	Y	Y	Y
Region FE	N	N	N	N	N	Y	Y	Y
Job-level FE	N	N	N	N	N	Y	Y	Y
Year FE	N	N	N	N	N	N	N	Y
Observations	38,021	28,926	28,925	28,925	27,910	27,910	27,910	27,910
R-squared	0.000	0.003	0.003	0.003	0.216	0.216	0.221	0.225

Notes. Robust standard errors clustered at the manager level in parentheses *** p<0.001, ** p<0.01, * p<0.05

Table 2. Self-Ratings in All Countries in Standard Years

	(1) Self-rating	(2) Self-rating	(3) Self-rating	(4) Self-rating	(5) Self-rating	(6) Self-rating	(7) Self-rating	(8) Self-rating
Female	-0.136***		-0.104***	-0.075***	-0.052***	-0.052***	-0.041**	-0.044**
	(0.009)		(0.011)	(0.014)	(0.015)	(0.015)	(0.015)	(0.015)
People of Color		-0.059***	-0.053***	-0.025	0.017	0.022	0.021	0.019
		(0.011)	(0.011)	(0.014)	(0.017)	(0.017)	(0.017)	(0.017)
Female*People				-0.068**	-0.072**	-0.075**	-0.072**	-0.071**
of Color				(0.021)	(0.024)	(0.024)	(0.024)	(0.024)
Constant	3.577***	3.567***	3.604***	3.594***	3.571***	3.569***	3.565***	3.567***
	(0.007)	(0.008)	(0.008)	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)
Manager FE	N	N	N	N	Y	Y	Y	Y
Region FE	N	N	N	N	N	Y	Y	Y
Job-level FE	N	N	N	N	N	Y	Y	Y
Year FE	N	N	N	N	N	N	N	Y
Observations	38,021	28,926	28,925	28,925	27,910	27,910	27,910	27,910
R-squared	0.008	0.001	0.006	0.006	0.241	0.242	0.246	0.249

Notes. Robust standard errors clustered at the manager level in parentheses *** p<0.001, ** p<0.01, * p<0.05

Table 3. Manager Ratings in All Countries in Standard and Non-Standard Year

	(1) Manager Rating	(2) Manager Rating	(3) Manager Rating	(4) Manager Rating	(5) Manager Rating	(6) Manager Rating
Female	0.021	0.019	0.021	0.035*	0.035*	0.040**
	(0.014)	(0.014)	(0.014)	(0.015)	(0.015)	(0.016)
People of Color	-0.054***	-0.054***	-0.056***	-0.047**	-0.052**	-0.060***
	(0.014)	(0.014)	(0.014)	(0.017)	(0.017)	(0.017)
Female*People of Color	-0.067**	-0.067**	-0.067**	-0.051*	-0.049*	-0.045*
	(0.020)	(0.020)	(0.020)	(0.023)	(0.023)	(0.023)
Non-Standard Year	-0.030**	-0.033**	-0.032**	-0.032*	-0.031*	-0.038**
	(0.009)	(0.011)	(0.012)	(0.013)	(0.013)	(0.013)
Non-Standard Year*Female		0.008		-0.003	-0.004	-0.003
		(0.018)		(0.018)	(0.018)	(0.018)
Non-Standard Year*People of Color			0.005	-0.005	-0.005	-0.001
			(0.018)	(0.018)	(0.018)	(0.018)
Constant	3.304***	3.275***	3.275***	3.268***	3.269***	3.271***
	(0.014)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)
Manager FE	N	N	N	Y	Y	Y
Region FE	N	N	N	N	Y	Y
Job-level FE	N	N	N	N	N	Y
Observations	37,813	37,813	37,813	36,952	36,943	36,943
R-squared	0.003	0.003	0.003	0.204	0.204	0.210

Notes. Robust standard errors clustered at the manager level in parentheses *** p<0.001, ** p<0.01, * p<0.05

Table 4. Counterfactual Simulation Varying Manager Ratings of Black Employees in the United States in Standard Years

Fraction of Black employees whose manager rating gets increased	Number of Black employees whose manager rating gets increased	Black employee	
		$\hat{\beta}_3$ (SE)	t-statistic *p*-value
0.20	172	-0.083 (0.035)	t = -2.314 p = 0.021
0.21	181	-0.072 (0.036)	t = -2.002 p = 0.045
0.22	**189**	**-0.063** **(0.036)**	***t* = -1.779** ***p* = 0.075**
0.23	198	-0.053 (0.036)	t = -1.481 p = 0.139
0.24	207	-0.043 (0.036)	t = -1.190 p = 0.234
0.25	215	-0.034 (0.036)	t = -0.933 p = 0.351
0.26	224	-0.024 (0.036)	t = -0.663 p = 0.507
0.27	232	-0.014 (0.036)	t = -0.392 p = 0.695
0.28	**241**	**-0.004** **(0.036)**	***t* = -0.117** ***p* = 0.907**
0.29	250	0.006 (0.036)	t = 0.180 p = 0.858

APPENDIX:

Table A.1. Unique Employee Data: Distribution by Gender and Race in All Countries (in %)

	Male	*Female*
All Employees	59.8%	40.2%
White Employees	45.3%	39.4%
Employees of Color	29.4%	34.8%
Global: Did not disclose race	25.3%	25.8%
	100%	100%
US: White	32.9%	52.7%
US: Black	21.4%	6.1%
US: Latinx	18.4%	5.4%
US: Asian	24.3%	29.0%
US: Other races	1.4%	3.1%
US: Did not disclose race	1.4%	3.8%
	100%	100%

Notes. This table is a cumulative summary of unique employee demographics across all four years (only including employees with non-missing manager ratings and self-ratings, which corresponds to 96% of the sample).

Table A.2.

Panel A: Average Self- and Manager Ratings by Demographic Group in All Countries in Standard Years

		Gender		Race		Gender-Race Interaction			
	All	Men	Women	Whites	People of color	White Men	White Women	Men of color	Women of color
Self-ratings	3.52 (0.76)	3.58 (0.76)	3.44 (0.75)	3.57 (0.74)	3.51 (0.77)	3.59 (0.75)	3.52 (0.73)	3.57 (0.78)	3.42 (0.76)
Manager ratings	3.22 (0.79)	3.24 (0.80)	3.20 (0.78)	3.28 (0.79)	3.20 (0.79)	3.27 (0.80)	3.30 (0.78)	3.22 (0.80)	3.17 (0.77)
Observations	38,022	23,347	14,674	17,103	11,823	10,976	6,127	6,812	5,010

Panel B: Average Self- and Manager Ratings by Demographic Group in All Countries in the Non-Standard Year

		Gender		Race		Gender-Race Interaction			
	All	Men	Women	Whites	People of color	White Men	White Women	Men of color	Women of color
Self-ratings	3.52 (0.75)	3.57 (0.76)	3.43 (0.75)	3.55 (0.74)	3.51 (0.76)	3.58 (0.74)	3.50 (0.72)	3.56 (0.77)	3.43 (0.75)
Manager ratings	3.19 (0.81)	3.20 (0.82)	3.18 (0.80)	3.25 (0.81)	3.17 (0.81)	3.24 (0.82)	3.26 (0.79)	3.19 (0.82)	3.16 (0.80)
Observations	11,664	7,279	4,385	5,256	3,532	3,477	1,879	2,067	1,465

Notes. Self-ratings refer to the self-evaluation that each employee has to fill out and share with their manager before the manager decides on their rating of the employee. The self-ratings row shows the average self-rating by each subgroup with standard deviations in parentheses. The manager ratings row shows the average rating that the corresponding subgroup receives from their managers (regardless of the managers' gender or race). The final row shows the total number of observations for each subgroup.

Table A.3. Manager Ratings Controlling for Manager Gender and Race for All Countries in Standard Years

VARIABLES	(1) Manager Rating	(2) Manager Rating	(3) Manager Rating	(4) Manager Rating	(5) Manager Rating	(6) Manager Rating	(7) Manager Rating
Female Employee	-0.005		-0.023*		-0.004	0.012	-0.004
	(0.010)		(0.011)		(0.013)	(0.011)	(0.013)
Female Manager	-0.046***		-0.074***		-0.081***		-0.081***
	(0.012)		(0.015)		(0.018)		(0.018)
Female Manager* Female Employee			0.061**		0.070**		0.070**
			(0.021)		(0.024)		(0.024)
Employee of Color		-0.074***		-0.074***	-0.079***	-0.075***	-0.077***
		(0.012)		(0.015)	(0.012)	(0.015)	(0.015)
Manager of Color		-0.014		-0.014		-0.014	-0.010
		(0.010)		(0.019)		(0.019)	(0.019)
Manager of Color* Employee of Color				0.003		0.003	0.001
				(0.025)		(0.025)	(0.025)
Constant	3.239***	3.299***	3.245***	3.285***	3.296***	3.281***	3.300***
	(0.007)	(0.016)	(0.007)	(0.009)	(0.010)	(0.010)	(0.011)
Manager FE	N	N	N	N	N	N	N
Job-level FE	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y
Observations	37,900	28,925	37,900	28,925	28,839	28,925	28,839
R-squared	0.013	0.012	0.014	0.012	0.014	0.013	0.014

Robust standard errors clustered at the manager level in parentheses *** p<0.001, ** p<0.01, * p<0.05

Table A.4. Manager Ratings Controlling for Self-Ratings in All Countries in Standard Years

	(1) Manager Rating	(2) Manager Rating	(3) Manager Rating	(4) Manager Rating	(5) Manager Rating	(6) Manager Rating	(7) Manager Rating	(8) Manager Rating
Female	0.034***		0.040***	0.057***	0.067***	0.067***	0.067***	0.063***
	(0.008)		(0.010)	(0.013)	(0.015)	(0.015)	(0.015)	(0.015)
People of Color		-0.055***	-0.057***	-0.041**	-0.060***	-0.068***	-0.073***	-0.076***
		(0.010)	(0.010)	(0.013)	(0.015)	(0.015)	(0.015)	(0.015)
Female*People of Color				-0.040*	-0.029	-0.025	-0.023	-0.021
				(0.019)	(0.022)	(0.022)	(0.022)	(0.022)
Self-rating	0.478***	0.471***	0.472***	0.472***	0.470***	0.471***	0.468***	0.466***
	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Constant	1.526***	1.604***	1.584***	1.579***	1.590***	1.589***	1.602***	1.611***
	(0.021)	(0.023)	(0.024)	(0.024)	(0.027)	(0.027)	(0.027)	(0.027)
Manager FE	N	N	N	N	Y	Y	Y	Y
Region FE	N	N	N	N	N	Y	Y	Y
Job-level FE	N	N	N	N	N	Y	Y	Y
Year FE	N	N	N	N	N	N	N	Y
Observations	38,021	28,926	28,925	28,925	27,910	27,910	27,910	27,910
R-squared	0.208	0.204	0.205	0.205	0.368	0.369	0.371	0.373

Notes. Robust standard errors clustered at the manager level in parentheses *** p<0.001, ** p<0.01, * p<0.05

Table A.5. Balance Check: Standard versus Non-Standard Year for All Countries

Sample characteristic	Standard years	Non-standard year	p-value
% Female	38.6%	37.6%	0.05
% People of Color	40.9%	39.7%	0.06
Self-rating	3.52	3.52	0.24

Table A.6. Manager Ratings in All Countries in Standard and Non-Standard Year with Self-Ratings and Lagged Manager Ratings

VARIABLES	(1) Manager Rating	(2) Manager Rating	(3) Manager Rating	(4) Manager Rating	(5) Manager Rating
Female	0.061***	0.061***	0.042**	0.042**	0.053***
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
People of Color	-0.068***	-0.069***	-0.034*	-0.034*	-0.051***
	(0.014)	(0.014)	(0.016)	(0.016)	(0.015)
Female*People of Color	-0.013	-0.013	-0.029	-0.029	-0.003
	(0.020)	(0.020)	(0.022)	(0.022)	(0.021)
Self-Ratings	0.459***	0.466***			0.410***
	(0.007)	(0.007)			(0.008)
Non-Standard Year	-0.029***	0.072	0.024*	-0.201***	-0.003
	(0.009)	(0.044)	(0.010)	(0.045)	(0.055)
Non-Standard Year*Self-Rating		-0.029*			-0.058***
		(0.012)			(0.013)
Lagged Manager Rating			0.320***	0.306***	0.227***
			(0.006)	(0.007)	(0.006)
Non-Standard Year*Lagged Manager Rating				0.068***	0.071***
				(0.013)	(0.013)
Constant	1.632***	1.609***	2.241***	2.291***	1.081***
	(0.025)	(0.026)	(0.023)	(0.025)	(0.032)
Manager FE	Y	Y	Y	Y	Y
Job-level FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Observations	36,943	36,943	28,864	28,864	28,864
R-squared	0.354	0.355	0.311	0.312	0.407

Notes. Robust standard errors clustered at the manager level in parentheses *** p<0.001, ** p<0.01, * p<0.05

Table A.7. Manager Ratings of Newcomers in All Countries

Panel A: Standard Years

	(1) Manager Rating	(2) Manager Rating	(3) Manager Rating	(4) Manager Rating	(5) Manager Rating	(6) Manager Rating	(7) Manager Rating	(8) Manager Rating
Female	-0.011		-0.001	0.010	0.055	0.054	0.044	0.043
	(0.019)		(0.023)	(0.034)	(0.038)	(0.038)	(0.038)	(0.038)
People of color		-0.124***	-0.124***	-0.114***	-0.101**	-0.098*	-0.099*	-0.103**
		(0.024)	(0.024)	(0.031)	(0.039)	(0.039)	(0.039)	(0.039)
Female*People of Color				-0.022	-0.034	-0.033	-0.039	-0.039
				(0.042)	(0.052)	(0.053)	(0.052)	(0.052)
Constant	2.836***	2.919***	2.919***	2.915***	2.902***	2.901***	2.907***	2.909***
	(0.015)	(0.022)	(0.024)	(0.027)	(0.021)	(0.021)	(0.021)	(0.021)
Observations	7,162	5,080	5,079	5,079	3,493	3,493	3,493	3,493
Manager FE	N	N	N	N	Y	Y	Y	Y
Region FE	N	N	N	N	N	Y	Y	Y
Job-level FE	N	N	N	N	N	N	Y	Y
Year-level FE	N	N	N	N	N	N	N	Y
R-squared	0.000	0.007	0.007	0.007	0.523	0.523	0.530	0.531

Panel B: Non-Standard Year

	(1) Manager Rating	(2) Manager Rating	(3) Manager Rating	(4) Manager Rating	(5) Manager Rating	(6) Manager Rating	(7) Manager Rating	(8) Manager Rating
Female	0.059		0.100*	0.073	-0.099	-0.104	-0.095	-0.095
	(0.034)		(0.040)	(0.056)	(0.083)	(0.085)	(0.091)	(0.091)
People of color		-0.060	-0.064	-0.092	-0.235*	-0.215*	-0.200*	-0.200*
		(0.042)	(0.041)	(0.055)	(0.094)	(0.101)	(0.100)	(0.100)
Female*People of Color				0.058	0.332*	0.315*	0.298*	0.298*
				(0.083)	(0.138)	(0.142)	(0.147)	(0.147)
Constant	2.712***	2.824***	2.778***	2.790***	2.902***	2.908***	2.901***	2.901***
	(0.025)	(0.033)	(0.036)	(0.041)	(0.048)	(0.051)	(0.053)	(0.053)
Observations	2,121	1,532	1,532	1,532	798	787	787	787
Manager FE	N	N	N	N	Y	Y	Y	Y
Region FE	N	N	N	N	N	Y	Y	Y
Job-level FE	N	N	N	N	N	N	Y	Y
Year-level FE	N	N	N	N	N	N	N	Y
R-squared	0.001	0.002	0.006	0.006	0.555	0.555	0.559	0.559

Table A.8. Manager and Self-Ratings in the United States and Other Countries in Standard Years

	Other countries			USA		
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Manager Rating	Self-Rating	Manager Rating	Manager Rating	Self-Rating	Manager Rating
Female	0.033	-0.074*	0.071*	0.043*	-0.037*	0.060***
	(0.034)	(0.031)	(0.029)	(0.019)	(0.018)	(0.018)
People of Color	-0.004	0.055	-0.032	-0.089***	0.001	-0.090***
	(0.035)	(0.034)	(0.030)	(0.021)	(0.020)	(0.018)
Female*People of Color	-0.084	-0.096*	-0.036	-0.041	-0.055	-0.017
	(0.046)	(0.043)	(0.040)	(0.030)	(0.029)	(0.027)
Self-Rating			0.506***			0.448***
			(0.013)			(0.009)
Constant	3.248***	3.479***	1.488***	3.279***	3.610***	1.664***
	(0.019)	(0.018)	(0.047)	(0.009)	(0.009)	(0.034)
Manager FE	Y	Y	Y	Y	Y	Y
Job-level FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Observations	8,800	8,800	8,800	18,749	18,749	18,749
R-squared	0.284	0.322	0.446	0.226	0.238	0.362

Notes. Robust standard errors clustered at the manager level in parentheses *** p<0.001, ** p<0.01, * p<0.05

Table A.9.

Panel A. Average Self- and Manager Ratings by Demographic Group in the United States in Standard Years

		Gender		Race				
	All employees	Men	Women	White	Black	Latinx	Asian	Other
Self-ratings	3.59 (0.75)	3.62 (0.76)	3.53 (0.74)	3.61 (0.74)	3.41 (0.82)	3.57 (0.78)	3.56 (0.75)	3.61 (0.80)
Manager ratings	3.25 (0.75)	3.25 (0.80)	3.25 (0.78)	3.29 (0.80)	2.99 (0.77)	3.16 (0.78)	3.23 (0.78)	3.19 (0.80)
Observations	19,977	12,321	7,656	12,258	861	849	4,906	491

Panel B. Average Self- and Manager Ratings by Demographic Group in the United States in the Non-Standard Year

		Gender		Race				
	All employees	Men	Women	White	Black	Latinx	Asian	Other
Self-ratings	3.58 (0.75)	3.61 (0.75)	3.53 (0.74)	3.61 (0.73)	3.42 (0.81)	3.51 (0.76)	3.56 (0.76)	3.54 (0.75)
Manager ratings	3.23 (0.81)	3.23 (0.82)	3.23 (0.80)	3.26 (0.81)	2.96 (0.80)	3.14 (0.81)	3.20 (0.81)	3.23 (0.79)
Observations	6,168	3,842	2,326	3,849	240	251	1,487	134

Notes. The self-ratings row shows the average self-rating by each subgroup with standard deviations in parentheses. The manager ratings row shows the average rating that the corresponding subgroup receives from their managers (regardless of the managers' gender or race). The final row shows the total sample size for each subgroup.

Table A.10

Panel A. Average Self- and Manager Ratings by Gender x Race Interaction in the United States in Standard Years

	White Men	White Women	Black Men	Black Women	Latinx Men	Latinx Women	Asian Men	Asian Women	Other Men	Other Women
Self-ratings	3.62 (0.74)	3.58 (0.72)	3.42 (0.82)	3.41 (0.83)	3.59 (0.78)	3.55 (0.77)	3.62 (0.77)	3.48 (0.72)	3.70 (0.81)	3.51 (0.78)
Manager ratings	3.28 (0.80)	3.32 (0.78)	2.97 (0.81)	3.00 (0.73)	3.13 (0.78)	3.20 (0.78)	3.26 (0.79)	3.21 (0.77)	3.22 (0.83)	3.14 (0.77)
Observations	8,032	4,226	404	457	445	404	2,825	2,081	271	220

Panel B. Average Self- and Manager Ratings by Gender x Race Interaction in the United States in the Non-Standard Year

	White Men	White Women	Black Men	Black Women	Latinx Men	Latinx Women	Asian Men	Asian Women	Other Men	Other Women
Self-ratings	3.62 (0.74)	3.57 (0.73)	3.41 (0.78)	3.43 (0.84)	3.45 (0.78)	3.59 (0.77)	3.63 (0.77)	3.47 (0.73)	3.61 (0.78)	3.48 (0.72)
Manager ratings	3.24 (0.81)	3.29 (0.80)	2.94 (0.83)	2.97 (0.77)	3.04 (0.80)	3.24 (0.81)	3.22 (0.81)	3.17 (0.81)	3.27 (0.86)	3.19 (0.71)
Observations	2,544	1,305	106	134	136	115	870	617	71	63

Notes. Self-ratings refer to the self-evaluation that each employee has to fill out and share with their manager before the manager decides on their rating of the employee. The self-ratings row shows the average self-rating by each subgroup on the left with standard deviations in parentheses. The manager ratings row shows the average rating that the corresponding subgroup receives from their managers (regardless of the managers' gender or ethnicity). The final row shows the total sample size for each subgroup.

Table A.11. Manager Ratings in the United States in Standard Years

	(1) Manager Rating	(2) Manager Rating	(3) Manager Rating	(4) Manager Rating
Female	0.001		0.005	0.036
	(0.011)		(0.011)	(0.019)
Asian		-0.062***	-0.063***	-0.054*
		(0.018)	(0.018)	(0.023)
Black		-0.279***	-0.280***	-0.278***
		(0.033)	(0.033)	(0.049)
Latinx		-0.097**	-0.097**	-0.080
		(0.034)	(0.034)	(0.048)
Other		-0.120**	-0.121**	-0.065
		(0.042)	(0.042)	(0.055)
Female*Asian				-0.028
				(0.033)
Female*Black				-0.011
				(0.065)
Female*Latinx				-0.043
				(0.067)
Female*Other				-0.135
				(0.082)
Constant	3.226***	3.259***	3.257***	3.248***
	(0.004)	(0.011)	(0.012)	(0.013)
Manager FE	Y	Y	Y	Y
Job-level FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Observations	37,091	37,091	37,091	37,091
R-squared	0.208	0.210	0.210	0.211

Notes. Robust standard errors clustered at the manager level in parentheses *** p<0.001, ** p<0.01, * p<0.05

Table A.12. Self-Ratings in the United States in Standard Years

VARIABLES	(1) Self-Rating	(2) Self-Rating	(3) Self-Rating	(4) Self-Rating
Female	-0.091***		-0.092***	-0.034
	(0.010)		(0.010)	(0.018)
Asian		-0.019	-0.008	0.019
		(0.018)	(0.018)	(0.022)
Black		-0.136***	-0.130***	-0.180***
		(0.036)	(0.037)	(0.053)
Latinx		0.021	0.023	0.029
		(0.033)	(0.033)	(0.045)
Other		0.037	0.042	0.103
		(0.043)	(0.043)	(0.057)
Female*Asian				-0.076*
				(0.031)
Female*Black				0.084
				(0.074)
Female*Latinx				-0.023
				(0.067)
Female*Other				-0.152
				(0.088)
Constant	3.562***	3.571***	3.606***	3.588***
	(0.004)	(0.011)	(0.012)	(0.012)
Manager FE	Y	Y	Y	Y
Job-level FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Observations	37,091	37,091	37,091	37,091
R-squared	0.237	0.236	0.238	0.239

Notes. Robust standard errors clustered at the manager level in parentheses *** p<0.001, ** p<0.01, * p<0.05

Table A.13. Manager Ratings in the United States Controlling for Self-ratings in Standard Years

VARIABLES	(1) Manager Rating	(2) Manager Rating	(3) Manager Rating	(4) Manager Rating
Female	0.043***		0.048***	0.052**
	(0.010)		(0.010)	(0.017)
Asian		-0.053***	-0.059***	-0.063**
		(0.016)	(0.016)	(0.020)
Black		-0.216***	-0.218***	-0.194***
		(0.030)	(0.030)	(0.043)
Latinx		-0.107***	-0.108***	-0.094*
		(0.031)	(0.031)	(0.044)
Other		-0.138***	-0.141***	-0.113*
		(0.037)	(0.037)	(0.050)
Female*Asian				0.007
				(0.031)
Female*Black				-0.050
				(0.057)
Female*Latinx				-0.032
				(0.061)
Female*Other				-0.063
				(0.076)
Self-ratings	0.472***	0.470***	0.472***	0.472***
	(0.006)	(0.006)	(0.006)	(0.006)
Constant	1.543***	1.579***	1.556***	1.555***
	(0.023)	(0.025)	(0.026)	(0.026)
Manager FE	Y	Y	Y	Y
Job-level FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Observations	37,091	37,091	37,091	37,091
R-squared	0.363	0.365	0.365	0.365

Notes. Robust standard errors clustered at the manager level in parentheses *** p<0.001, ** p<0.01, * p<0.05

Table A.14. Counterfactual Simulation Varying Manager Ratings of Asian Employees in the United States in Standard Years

Fraction of Asian employees whose manager rating gets increased	Number of Asian employees whose manager rating gets increased	Asian employee	
		$\hat{\beta}_3$ (SE)	t-statistic *p*-value
0.01	49	-0.059 (0.019)	*t* = -3.173 *p* = 0.002
0.02	98	-0.050 (0.019)	*t* = -2.655 *p* = 0.008
0.03	147	-0.040 (0.019)	*t* = -2.157 *p* = 0.031
0.04	**196**	**-0.031** **(0.019)**	***t* = -1.626** ***p* = 0.104**
0.05	245	-0.21 (0.019)	*t* = -1.103 *p* = 0.270
0.06	294	-0.12 (0.019)	*t* = -0.612 *p* = 0.540
0.07	**343**	**-0.002** **(0.019)**	***t* = -0.114** ***p* = 0.909**
0.08	392	0.007 (0.019)	*t* = 0.364 *p* = 0.716

Table A.15. Counterfactual Simulation Varying Manager Ratings of Latinx employees in the United States in Standard Years

Fraction of Latinx employees whose manager rating gets increased	Number of Latinx employees whose manager rating gets increased	Latinx employee	
		$\hat{\beta}_3$ (SE)	t-statistic *p*-value
0.01	8	-0.093 (0.035)	*t* = -2.656 *p* = 0.008
0.02	17	-0.083 (0.035)	*t* = -2.358 *p* = 0.018
0.03	25	-0.074 (0.035)	*t* = -2.090 *p* = 0.037
0.04	**34**	**-0.064** **(0.036)**	***t* = -1.802** ***p* = 0.072**
0.05	42	-0.055 (0.036)	*t* = -1.542 *p* = 0.123
0.06	51	-0.045 (0.036)	*t* = -1.266 *p* = 0.206
0.07	59	-0.036 (0.036)	*t* = -0.992 *p* = 0.321
0.08	68	-0.026 (0.036)	*t* = -0.713 *p* = 0.476
0.09	76	-0.016 (0.036)	*t* = -0.455 *p* = 0.649
0.10	85	-0.006 (0.036)	*t* = -0.158 *p* = 0.875
0.11	**93**	**0.003** **(0.036)**	***t* = 0.082** ***p* = 0.934**

Table A.16. Counterfactual Simulation Varying Manager Ratings of "Other" Employees in the United States in Standard Years

Fraction of "Other" employees whose manager rating gets increased	Number of "Other" employees whose manager rating gets increased	Employee with self-selected "Other" racial category	
		$\widehat{\beta}_3$ (SE)	t-statistic *p*-value
0.01	5	-0.101 (0.043)	*t* = -2.348 *p* = 0.019
0.02	10	-0.092 (0.043)	*t* = -2.107 *p* = 0.035
0.03	**15**	**-0.082** **(0.044)**	***t* = -1.877** ***p* = 0.061**
0.04	20	-0.072 (0.044)	*t* = -1.640 *p* = 0.101
0.05	25	-0.062 (0.044)	*t* = -1.418 *p* = 0.156
0.06	29	-0.054 (0.044)	*t* = -1.230 *p* = 0.219
0.07	34	-0.045 (0.044)	*t* = -1.025 *p* = 0.306
0.08	39	-0.035 (0.044)	*t* = -0.789 *p* = 0.430
0.09	44	-0.026 (0.045)	*t* = -0.573 *p* = 0.567
0.10	49	-0.015 (0.045)	*t* = -0.345 *p* = 0.730
0.11	54	-0.006 (0.045)	*t* = -0.139 *p* = 0.889
0.12	**59**	**0.003** **(0.045)**	***t* = 0.077** ***p* = 0.939**