

You're too kind: How employee self-ratings and evaluator gender jointly affect subjective performance ratings

Andrew H. Newman*
University of South Carolina
Andrew.Newman@moore.sc.edu

Jason T. Rasso
University of South Carolina
Jason.Rasso@moore.sc.edu

Bryan R. Stikeleather
University of South Carolina
Bryan.Stikeleather@moore.sc.edu

Acknowledgements: We thank Kai Bauch (discussant) and participants at the EIASM Conference on Performance Measurement and Management Control for their helpful feedback. We also thank Dustin McDowell for providing research assistance and gratefully acknowledge research support from the Riegel and Emory Human Resource Center and the Darla Moore School of Business at the University of South Carolina. The authors have no conflicts of interest to disclose.

Data Availability: Data are available upon request

JEL Classifications: M12, M14, M41, M51, M54

Key words: performance ratings; self-assessment; gender; leniency

You're too kind: How employee self-ratings and evaluator gender jointly affect subjective performance ratings

ABSTRACT

Research suggests evaluators' numerical performance ratings of employees tend to exhibit leniency, which can significantly diminish the value of performance evaluations. We investigate whether receiving employees' numerical self-ratings of their performance leads evaluators to provide higher performance ratings than they would otherwise and whether this effect is more pronounced among male versus female evaluators. Using an experiment, we find that, consistent with our theoretical predictions, male evaluators provide lower numerical ratings than female evaluators, providing employees' self-ratings to evaluators significantly increases their ratings, and this effect is more pronounced among male than female evaluators. Also, evaluators forming a preliminary rating before receiving an employee's self-rating does *not* mitigate this effect. By providing insight into how a key component of the evaluation process, employee self-ratings, can interact with evaluator gender to affect leniency, we highlight the potential for the objectivity and usefulness of performance evaluations being compromised, particularly among male evaluators.

I. INTRODUCTION

This study investigates whether receiving employees' numerical performance self-ratings leads evaluators to act more leniently when rating employee performance and whether this effect is more pronounced among male versus female evaluators. We define evaluators as acting *more leniently* than other evaluators when they provide relatively higher performance ratings given the same objective performance measures. In practice, evaluators who act more leniently tend to cluster employee ratings at the upper end of the performance rating scale, thereby decreasing the extent to which the ratings discriminate between high- and low-performing employees (Moers, 2005; Bol, 2011; Golman and Bhatia 2012; Cheng et al. 2017).

Due to the noise inherent in many performance measures in practice and given the general infeasibility of developing complete employment contracts that account for all possible scenarios, firms often rely on evaluators to use subjective performance ratings to incorporate relevant information such as unforeseen or uncontrollable events that might impact objective performance measures (Baker et al. 1994; Gibbs et al. 2004; Wick 2021). Doing so can increase how fair employees feel about the appraisal process while also enabling the performance evaluation process to more readily and effectively identify, promote, reward, and develop high-performing employees, train and demote (or layoff) low-performers, and appropriately compensate and motivate employees in general (Bonner and Sprinkle 2002; Kelly et al. 2015; Bol 2008). Therefore, by reducing the differentiation in performance between higher- and lower-performing employees, rating leniency diminishes the value of performance ratings to the firm and can undermine the integrity of performance management systems. Accordingly, understanding the factors that affect evaluation leniency is essential for creating fairer and more effective appraisal systems. Our study investigates whether a common feature of the performance rating process, namely eliciting employee self-ratings, fosters more lenient ratings from

evaluators. Further, prior research finds that gender significantly influences decisions and behavior in organizational settings (e.g., Lyness and Heilman 2006; Roberson and Kulik 2007; Bowles and Gelfand 2010; Heilman 2012; Rice and Barth 2016; Chandler 2018). Given this, and the fact that males continue to occupy the majority of jobs with evaluation responsibility (Catalyst 2023; Fuhrmans 2024), our study's primary focus is to investigate whether the effect of eliciting employee self-ratings interacts with evaluator gender to influence leniency.

In practice, evaluators often seek subjective input from employees when rating their performance. This has been shown to provide multiple benefits to the firm, such as promoting perceptions of procedural fairness in decision-making, improving morale, increasing accountability, boosting motivation, aligning beliefs, and building trust between evaluators and their employees (Fulk et al. 1985; Meyer 1991; Folger et al. 1992; Morrison and Milliken 2000; Erdogan et al. 2001; Hoogervorst et al. 2013). As discussed further in the next section, one type of quantitative employee input that evaluators often elicit during the rating process is an employee's numerical rating of the employee's own performance (i.e., a self-rating) using the same rating scale that evaluators will use to assess the employee. Evaluators can then consider such numerical self-ratings when evaluating the employee.

We first leverage gender role theory (Bem 1981; Bussey and Bandura 1999; Eagly and Wood 2012) to predict that male evaluators will be more strict than female evaluators in their subjective performance ratings. We next couple theory and research that finds individuals assess their own performance as higher than deserved as they often believe they perform better than they actually do (Kruger and Dunning 1999; Judge and Bono 2001; Dunning et al., 2003 2004; Sedikides and Gregg 2008; Kromrei 2015; Cristofaro and Giardino 2020) with theory and research on anchoring and adjustment (Tversky and Kahneman 1974; Furnham and Boo 2011) to

predict a positive main effect of receiving employee self-ratings on evaluator leniency. That is, providing evaluators with employees' generally inflated self-ratings along with objective performance measures will lead evaluators to provide higher performance ratings relative to evaluators who only receive objective performance measures and do not receive employees' self-ratings. Finally, and the primary focus of our study, we develop a more comprehensive theory that integrates gender role theory and anchoring and adjustment theory to predict an interaction in which the tendency for employees' self-ratings to result in more lenient evaluator performance ratings will be more pronounced among male evaluators than among female evaluators.

We use a 3 (*Rating Setting*) \times 2 (*Male Identification*) \times 2 (*Preliminary Rating*) experiment to test our predictions. First, we run an independent study in which participants assume the role of employees, perform a real-effort task where their performance output is subject to an uncontrollable event that either increases or decreases their productivity *ex post*, and then, after learning about their adjusted performance level, they self-rate their performance on a numerical ratings scale. Second, we conduct our primary study in which a different group of participants assume the role of evaluators, observe only an employee participant's adjusted task performance (but do know the nature of any adjustment), and are tasked with providing a numerical rating of that participant's performance. Thus, our setting reflects one whereby evaluators make subjective assessments in the face of noisy performance information while utilizing a similar experimental approach used by prior research in this area (e.g., Bol and Smith 2011; Karagozoglu and Riedl 2015; Bol et al. 2016; Luft et al. 2016; Demere et al. 2019).

Within this context, we operationalize three types of rating settings. In the first setting, evaluators receive performance data but not the employee's self-rating, and then provide their rating of the employee's performance. In the second setting, evaluators receive an employee's

performance data and observe the employee's self-rating, then provide their numerical rating of the employee's performance. The third setting replicates the second, except that evaluators also receive a narrative statement justifying the employee's self-rating. Employees often provide narrative statements to accompany and justify their numerical self-ratings in practice; thus, the third setting allows us to examine *ex post* the potential importance of the inclusion of narrative statements with respect to the influence of self-ratings on evaluators' rating decisions.

We also recognize that evaluators may have preconceived notions about employee performance prior to receiving an employee's self-rating. In order to reflect this as well as be able to examine *ex post* the potential effect it might have with respect to the influence of self-ratings on evaluators' rating decisions, we also manipulate whether evaluators form a preliminary rating of the employee's performance prior to receiving the employee's self-rating and making their final rating. Finally, across all conditions, we collect self-reported demographic data, including participants' primary gender identity.

Our analysis indicates that the second rating setting (i.e., receiving a self-rating without accompanying justification) and the third rating setting (i.e., receiving a self-rating accompanied by justification) yielded statistically similar ratings. That is, providing evaluators with a narrative justification for the self-rating did not affect the mean ratings they gave relative to providing the self-rating without justification; thus, we pooled these two settings together when testing our hypotheses. Our results are consistent with our predicted interaction. Evaluators provided higher performance ratings when they received employees' self-ratings relative to when they did not, and this effect was more pronounced among males than females. Indeed, providing self-ratings led to significantly higher ratings among males but did not among females. Notably, although males were more susceptible to employees' self-ratings than females, we also found that male

evaluators generally gave lower performance ratings than female evaluators across all conditions. Finally, we find no empirical evidence that having evaluators form their own preliminary rating of employee performance before receiving the employee's self-rating alters the influence of employees' self-ratings on evaluators' ratings. This null result suggests that the phenomena we document are robust to settings in which evaluators already have preconceived notions about the numerical rating an employee deserves.

Acting more leniently when assigning performance ratings reduces the value of those ratings for firms. Our study contributes to prior research on the potential causes and effects of evaluation leniency (e.g., Farh and Dobbins 1989; Moers 2005; Bol 2011; Golman and Bhatia 2012; Cheng et al. 2017). We do so by examining a rather understudied yet practically relevant setting in which managers make evaluation decisions in the face of performance information noise and investigating how two elements of the performance evaluation process commonly found in practice, yet previously unstudied, could jointly drive evaluators to act more leniently in such a setting. Specifically, we develop and test theory regarding how a contextual feature commonly found in evaluation processes, namely the provision of employees' numerical self-ratings to their evaluator as part of the evaluation process, and a demographic characteristic of evaluators, namely their gender identification, *interact* to influence how leniently evaluators subjectively evaluate employee performance when there is noisy performance information.

While providing important insights into what may drive leniency in performance evaluations, our findings also underscore a challenge for firms. Specifically, although eliciting employee involvement and engagement with the performance rating process can benefit firms in multiple potential ways, our study highlights that one common approach of doing so could potentially lead to less useful performance ratings. Thus, our study suggests that firms should

carefully consider how any benefits associated with allowing employees to numerically self-rate their performance compare to the potential costs of diminishing the utility of this evaluation process. Moreover, we develop theory for why males and females will handle uncertainty in performance information differently and how that will affect the information they incorporate into their evaluation decisions. In doing so, our study highlights that gender matters when considering how evaluators might incorporate employee-generated performance rating information. The relevance of gender in our findings, in particular that male evaluators are more likely than female evaluators to demonstrate leniency influenced by employees' self-ratings, is underscored by the fact that male evaluators have historically comprised a disproportionate share of management (and continue to do so) (Eagly and Carli 2007; Grossman et al. 2019; Catalyst 2023; Fuhrmans 2024).

II. BACKGROUND AND THEORY DEVELOPMENT

The prevalence of employee self-ratings in practice

Our study presumes that firms' performance rating processes often incorporate employees' numerical self-ratings as a formal input used by evaluators. To provide empirical evidence and insights into the prevalence of this phenomenon in practice, we conducted a brief survey and collected 105 responses from currently employed participants on the CloudResearch Connect platform. Survey participants received \$0.50 for answering a series of questions about whether they had received a recent performance evaluation and, if so, various aspects of the evaluation process. We tabulate and report the data in the Appendix.

We find 83.8 percent of respondents (88/105) indicated they had received a performance evaluation in the last three years, and 61.4% (54/88) of these indicated that they were asked to provide a performance self-assessment as part of the process. Of the 54 individuals who

indicated they were asked to self-assess their own performance, 75.9% (41 participants) were asked to do so using a numerical ratings scale. Of the 41 participants who provided a numerical self-rating, 98% (40/41) indicated they used the same scale as their evaluator.

Also, 47 of the 54 (87%) participants who were asked to self-assess their own performance indicated that they were asked to justify their self-assessment, while only 29.6% were asked to provide other materials beyond a self-rating or justification for the self-rating to support their self-assessment. Overall, the survey results provide empirical evidence consistent with a key presumption of our study, namely that employees providing self-ratings is a common feature of the performance ratings process.

Leniency in performance ratings

Despite their importance, performance ratings are often subject to biases and subjective influences. For example, research finds performance ratings tend to exhibit leniency, i.e., a disproportionate number of employees are classified as "higher-performers" (Moers 2005; Bol 2011; Golman and Bhatia 2012; Cheng et al. 2017). This, in turn, leads to employee performance ratings that are clustered at the upper end of the performance spectrum. Leniency may reflect a bias among evaluators, but it does not necessarily have to do so. In our study, we are interested in comparing *relative levels* of rating leniency across different settings and *not* in detecting if there is a "leniency bias" per se.

Prior research documents that various factors can influence evaluator rating leniency. This research can be broadly categorized into four areas of inquiry, namely research on (i) the psychological and personality forces that foster it, (ii) the economic incentives and tradeoffs associated with it, (iii) demographic characteristics of those being evaluated that can affect it, and (iv) contextual features of the evaluation process itself that may promote it. Within the realm

of psychology and personality, evaluators' own personality traits (e.g., agreeableness, degree of extroversion, emotional stability) as well as their attitudes and beliefs regarding the rating process have been found to influence their degree of leniency (e.g., Tziner et al. 2005; Cheng et al. 2017; Jawahar 2001). From a more economics-based perspective, strategic elements can also influence evaluators' leniency, as Bol (2011) finds evidence that evaluators consider the tradeoff between the benefit of accurate evaluations versus the cost of gathering information and conducting analysis as well as their own incentives. Also, employees consider strategically engaging in impression management tactics to influence future evaluations (Wayne and Liden 1995).

Meanwhile, a variety of demographic characteristics and dynamics related to the race, gender, age, nationality, and sexual preference of *evaluatees* have been found to affect performance rating leniency (e.g., Heilman 1983; Kraiger and Ford 1985; Greenhaus et al. 1990; Ragins and Cornwell 2001; Lyness and Heilman 2006; Roberson and Kulik 2007; De Pater et al. 2010; King et al. 2010). Regarding contextual factors, elements such as an evaluation's format, social context (i.e., whether or not it is face-to-face), purpose (e.g., evaluative versus developmental), the uniqueness of the task or job being evaluated, team diversity, the evaluators' span of control as well as their general attitudes and beliefs towards the organization, and even the broader organizational culture can affect evaluator leniency (e.g., Farh et al. 1991; Tziner et al. 2005; Yun et al. 2005; Joshi and Roh 2009; Tan 2019; Gong et al. 2021).

Although research has investigated various factors that influence evaluator leniency as well as settings where leniency might be more likely to manifest, there is a scarcity of research related to factors influencing evaluator rating leniency in settings like ours where evaluators perform subjective ratings in the face of performance measure noise. Additionally, prior research

has not typically examined how evaluators' demographic characteristics influence their propensity to be lenient. As discussed above, most of the research on demographic characteristics examines how the demographic characteristics of *employees* affect evaluators' performance rating patterns. To help fill that void, our study focuses on whether and how a demographic characteristic of *evaluators*, namely their gender identification, affects their performance ratings leniency. Moreover, we are also interested in how a contextual feature commonly found in ratings processes, but understudied with regard to ratings leniency, namely the provision of employees' numerical self-ratings to their evaluators as part of the ratings process, affects evaluator leniency. We study these two relevant elements together because beyond their practical relevance, as detailed in the next section, we develop theory to predict that they will *interact* to influence performance ratings leniency, and, as discussed later, this interactive effect has important implications for practice.

Theory and hypotheses development

We begin this section by developing theory about the main effects of evaluator gender and employee self-ratings on evaluators' subjective rating of employee performance. Although these main effect predictions are not the primary focus of our study, developing these hypotheses provides important background information and insights that assist us with developing theory about how evaluator gender and employee self-ratings interact to affect evaluators' ratings.

Within our setting, in which evaluators provide subjective ratings in the presence of performance measure noise, we posit that evaluators' gender will influence their decisions. Gender role theory posits that males and females are socialized to exhibit different behaviors and attitudes (Bem 1981; Bussey and Bandura 1999; Eagly and Wood 2012). For example, males are traditionally more associated with traits such as assertiveness, competitiveness, and objectivity,

while females are more associated with empathy, nurturing, and interpersonal sensitivity (Hoffman 1977; Feingold 1994; Lippa 2005). Importantly, gender role theory and related research suggest that these socialized behaviors extend to professional settings, influencing differences in how males and females are evaluated by others (e.g., Ridgeway 2001; Eagly and Karau 2002; Lyness and Heilman 2006; Heilman 2012).

However, research has paid much less attention to the extent to which such socialized behaviors might influence differences in how males and females evaluate others (Rice and Barth 2016).¹ Gender role theory suggests that male evaluators, adhering to their social roles, will tend to prioritize task-oriented criteria and objective measures of performance when evaluating others' performance. In doing so, we expect male evaluators to focus relatively more on measurable outcomes and less on factors that are more difficult to measure (i.e., noisier) but that could have helped or hurt the employee's performance, as reflected in the objective measure. For example, male sales managers might be more likely to focus on objective measures such as a salesperson's number of sales, while downplaying the potential influence of less readily available or measurable variables such as weather- or economic-related effects.

Meanwhile, gender role theory suggests that female evaluators, in line with their social roles, will adopt a more empathetic approach to their evaluations. In doing so, we expect females

¹ Although not specifically related to evaluators subjective performance ratings, there is a stream of research examining how the alignment of evaluator-evaluatee gender affects evaluator decision-making more generally. For example, with respect to hiring decisions (often in education settings), research shows that evaluators (especially males) adhere to gender stereotypes while also preferring candidates of their own gender (e.g., Bosak and Sczesny 2011; Rice and Barth 2016; Chandler 2018). Also, Fanning et al. (2021) document gender-based differences in how male and female job candidates are evaluated during group recruiting events, finding that females (males) receive lower (higher) evaluations, particularly from male recruiters, during group recruiting events when they exhibit stereotypically male behaviors. In terms of how evaluators assess and punish employee workplace deviancy (e.g., covering up mistakes or lying about hours worked), Bowles and Gelfand (2010) find evaluator gender interacts with the gender of the worker who commits the deviant behavior, as females are more even-handed in their reactions while males are sticter toward females than males committing deviant behavior. The role of employee gender and how its alignment with evaluator gender influences evaluators' performance ratings is beyond the scope of our study; however, we discuss in our Conclusion the potential relevance for future research of such considerations.

to afford more consideration to the possible influence of potentially relevant yet noisier factors. More specifically, although female evaluators know that these factors can sometimes benefit employee performance, we expect they will be more attuned to the potential that such factors could also harm employee performance, such that the actual employee performance output available for evaluation may have been even greater if not for these factors. Given such uncertainty, we expect females to be more willing than their male counterparts to give employees "the benefit of the doubt." Taken together, we expect that male evaluators will rate a given level of employee performance less favorably than female evaluators. More formally, we predict:

H1: Male evaluators will provide lower employee performance ratings than female evaluators.

For H1, we developed theory on the influence of evaluators' gender on their relative performance ratings due to inherent differences in how males and females approach, interpret, and weight objective yet noisy employee performance information. For H2, we temporarily set aside gender as we now turn to anchoring theory and consider, more generally, the potential role that employees' subjectively determined numerical self-ratings could play in fostering leniency in evaluators' employee performance ratings.

In practice, firms expect (and research suggests) that eliciting employee self-assessments can have multiple benefits, such as enhancing employees' perceived fairness and comprehensiveness of evaluations, improving employees' personal development, and creating better alignment of perspectives and opinions via enhanced communication between supervisors and their employees (Harris and Schaubroeck 1988; Meyer 1991). Despite their potential benefits, self-enhancement theory suggests that individuals tend to present themselves in a favorable light, emphasizing their strengths and downplaying weaknesses (Sedikides and Gregg 2008). Thus, self-assessments are vulnerable to individuals' flattering beliefs about themselves,

which can be influenced by various factors, including the employee's self-esteem, cognitive biases, and/or lack of self-awareness (Atwater and Yammarino 1997; Fletcher 1999; Kruger and Dunning 1999; Dunning et al. 2003, 2004; Kromrei 2015). For instance, Dunning et al. (2004) reviewed research on employee self-assessment in healthcare and found that employees tend to overestimate their skills and display overconfidence in their judgments. Consistent with this theory and research, we expect that in our setting, employees' numerical self-ratings will tend to fall toward the upper end of the evaluation spectrum.

When evaluators receive employees' numerical self-ratings, they become subject to the employees' self-enhancement efforts before finalizing their performance assessment of the employee. We predict that when evaluators review employees' self-ratings, these ratings will serve as an anchor for evaluators' subsequent judgments, leading them to provide higher employee performance ratings relative to when their ratings are determined in the absence of the employees' self-flattering ratings.² The anchoring effect (Tversky and Kahneman 1974), whereby individuals rely heavily on an initial piece of information (the "anchor") when making decisions, explains this predicted behavior.

Research on anchoring (e.g., Epley and Gilovich 2001, 2006; Furnham and Boo 2011) studies how individuals create and adjust numerical estimates. For example, individuals may be asked to estimate the number of jelly beans in a jar. Exposing individuals to a numerical estimate before having them provide their own estimate causes them to "anchor" their estimate toward the numerical estimate they were first exposed to and then adjust from that. This effect leads to

² When significant information asymmetry exists between employees and their evaluators concerning factors that could potentially have affected employee performance, employees could attempt to also convey private information to the evaluator (outside of the self-rating) that present themselves in a favorable light. However, evaluators may discount the credibility of this information given it reflects "cheap talk." We abstract away from the complexity inherent in evaluatees having multiple channels with which to influence their performance rating by focusing on a setting in which employees cannot engage in any such informal communication outside of the ratings process or engage in any form of additional signaling to evaluators.

estimates that are tilted toward the original anchor, even if the anchor is randomly generated and non-diagnostic. Similarly, we expect that inflated numerical self-ratings will act as an anchor, leading evaluators to rate performance upward. Ultimately, even if evaluators are already predisposed to providing leniency in their ratings, we predict employees' self-ratings can validate, if not exacerbate, this tendency. Therefore, we make the following prediction:

H2: Evaluators will provide higher employee performance ratings when they receive an employee's self-rating prior to making their rating decision than when they do not receive it.

For our final hypotheses, we build on the underlying theory discussion for H1 and H2, to develop theory for why evaluators' gender will interact with the predicted anchoring effects associated with employees' self-ratings to influence evaluators' ratings. In other words, we develop theory to predict that the extent to which evaluators weight employees' self-ratings in their performance ratings depends on evaluator gender.

As discussed, employee self-ratings introduce a highly subjective (and likely upwardly biased) element into the ratings process. Regarding self-promotion and gender norms, males are generally more comfortable with and expected to engage in self-promotional behavior more frequently than females, which is often viewed as a positive trait in professional settings (Rudman 1998; Rudman and Glick 2001; Eagly and Karau 2002). Consequently, we expect male evaluators to be more likely to respond more favorably to self-ratings that appear to align with these norms, interpreting them as a sign of confidence, competence, and capability. In turn, we expect males to place greater weight on the confidence and assertiveness displayed in these self-ratings (i.e., place more weight on the employee-generated anchor), resulting in a higher rating relative to when they do not receive employees' self-ratings.

In contrast, female evaluators, given their personal gender lens that traditionally values modesty and humility (Eagly and Karau 2002), are more likely to perceive flattering self-promotion as overconfidence or even arrogance. In turn, we expect this perception to lead to a more skeptical evaluation of self-ratings, making female evaluators more likely to discount the legitimacy of the employee's ratings (and thus place less weight on the employee-generated anchor). This should result in female evaluators being less likely than their male counterparts to be influenced upward by self-ratings. Moreover, to the extent that female evaluators dismiss the self-ratings altogether then they could have no meaningful influence on their performance rating whereas to the extent that female evaluators view them as doing more "harm than good" then it could even result in lower evaluations relative to when female evaluators do not receive employees' self-ratings.

Consequently, we expect that the positive effect of employees' self-ratings on evaluators' performance ratings will be more pronounced among males than among females. We therefore make the following prediction:

H3: The positive effect of receiving an employee's self-rating on evaluator ratings will be stronger when evaluators are male versus female.

III. METHOD

Broad overview

We conduct an experiment using a 3 (*Rating Setting*) \times 2 (*Male Identification*) \times 2 (*Preliminary Rating*) design to test our hypotheses. In a stand-alone study, participants assume the role of Employees, perform a real-effort task, and self-rate their performance on a numerical rating scale. The task's performance measure is noisy in that it reflects not only the Employee's performance but also an uncontrollable shock to performance. Then, in the primary study, a new group of participants assumes the role of Evaluators, receives performance information

concerning an Employee and possibly the Employee's self-rating, and is then tasked with evaluating that Employee's performance using the same numerical rating scale.³

As detailed below, our experimental task tests our theory by having participants perform and evaluate an abstract task that does not require domain-specific knowledge, experience, or skill (Libby et al. 2002). Using an abstract task both fosters the generalizability of the findings across domains and is simple enough that individuals can quickly and easily learn its requirements and what must be done to perform well, thereby enabling evaluators to set reasonable performance expectations. Thus, we recruit from the available subject pool on CloudResearch Connect. To qualify to participate, participants had to be based in the U.S. and be at least 18 years old. Participants received \$2.00 in fixed pay for completing the task, plus a variable bonus based on the performance of the Employee assigned to them and their evaluation of that Employee (more details provided below).

Experimental procedures

Employee "seed data"

To provide performance data for Evaluators to assess, we used performance data previously collected as part of a separate experimental study, in which participants recruited via the Amazon Mechanical Turk platform served as Employees. These Employees performed 10 rounds of a real-effort task, each lasting 60 seconds. Specifically, they worked on an adapted version of the "slider" task described by Gill and Prowse (2015) and Chan (2017), among others. The task required participants to use their computer mouse to slide a dot to a designated spot on a line repeatedly for multiple periods. The designated spot varied with each subsequent slider, but the task itself remained the same. Each dot correctly positioned generated two "performance

³ The experimental design was approved by the authors' University's Institutional Review Board.

points" for the Employee. An illustration of the task is provided in Figure 1. In addition to effort, the level of performance points earned was also subject to uncontrollable, exogenous environmental shocks unrelated to effort. Specifically, these shocks increased or decreased the Employee's performance points total for a round by 10 percent (i.e., the performance measure could differ by 10 percent in either direction, depending on whether the Employee experienced a positive or negative shock). Participants learned of the shock at the end of each round when they received feedback regarding the profit points they had earned for that round.

[Figure 1]

For each round, employees were asked to self-rate their performance on an 11-point numerical rating scale ranging from 0 (poor performance) to 10 (outstanding performance). Specifically, Employees were asked to complete the following statement: "*I believe that I deserve a rating of ____.*" Employees were told that their performance would be evaluated by Evaluators, who would observe their performance points but would not know if they experienced a positive or a negative shock to their performance. Thus, evaluators would not know employees' true performance output. They were also told that their self-rating might or might not be provided to Evaluators.

Using previously acquired employee-level data enables us to provide Evaluators with realistic performance data and self-ratings cost-effectively, rather than arbitrarily setting these parameters ourselves. We selected a representative sample of the performance and self-ratings data for use in our primary study. In cases where multiple Employees had the same level of performance points but differed in their self-ratings, we used the average self-rating as a proxy for how employees viewed their performance. Note that we are not directly interested in

Employee behavior in this study and use Employee output only to provide "seed data" for Evaluators to evaluate. We now discuss in detail the procedures related to the Evaluators.

Evaluators

We assigned Evaluators to one of six conditions ($3 \text{ Rating Setting} \times 2 \text{ Preliminary Rating}$). We described the task that Employees had to complete to Evaluators and informed them that a performance round for Employees lasted 60 seconds and that the performance points Employees earned were influenced by their performance (two points per successful slider) as well as uncontrollable shocks of (+ / -) 10 percent. Evaluators were also allowed to practice the same task that Employees performed (also for 60 seconds). This allowed them to become familiar with the task requirements and to develop a sense of what might constitute reasonable performance outcomes. Each Evaluator was responsible for providing a single performance evaluation of one assigned Employee for one performance round.

Every Employee was evaluated by six independent Evaluators, one in each condition. This allowed us to control for any effect that variation in employee performance and/or self-ratings might have across our experimental conditions. All six Evaluators first received just the Employee's performance points. Evaluators were assigned to one of two preliminary rating (*PR*) conditions. In the *PR* condition, evaluators were asked to provide a preliminary rating about the employee's performance using an 11-point ratings scale bounded by 1 (Poor performance) and 10 (Outstanding performance), i.e., the same scale used by Employees to self-rate their performance, prior to providing their final performance rating. Any such preliminary ratings occur *after* receiving *objective* performance information concerning the Employee but *prior* to receiving an Employee's self-rating (discussed next). Although not part of our underlying theory development, we acknowledge that evaluators may very well form preliminary assessments

before receiving any self-rating information from employees.⁴ Therefore, capturing any effects this might have on the ultimate impact of employees' self-ratings on evaluator ratings allows us to control for its impact with respect to our variables of interest. In the *No PR* condition, evaluators were not asked to provide a preliminary rating; instead, they proceeded with the experiment.

Next, regardless of *PR* condition, Evaluators were assigned to one of three *Rating Setting* conditions (*No SR*, *SR*, and *SR+*). In the *No SR* condition, Evaluators did not receive any information about the Employee's self-rating before they provided a final rating of the Employee's performance using the 11-point ratings scale previously described.⁵ In the *SR* condition, prior to making a final rating, Evaluators received the Employee's self-rating and then provided a final rating. Specifically, Evaluators were informed that the Employee self-assessed their performance using the same ratings scale used by the Evaluator and indicated, "*I believe that I deserve a rating of ____.*" In the *SR+* condition, evaluators received the Employee's self-rating statement and a brief statement justifying the self-rating before providing their final rating. Specifically, following the Employee's self-rating statement, Evaluators were told that their Employee provided the following message: "*As you know, there were factors beyond my control that affected my points total. I deserve a rating of ____ because of the effort I put into this task.*"

For experimental control purposes, all Evaluator participants in the *SR+* condition received the same qualitative statement. This allowed us to investigate the potential influence of a qualitative statement while also providing a justification universally applicable to all

⁴ Whether the likelihood of such preliminary assessments are due to underlying human nature, influenced by Evaluator-specific characteristics, and/or affected by Evaluators' relationship with the Employee are beyond our scope of interest in this study.

⁵ Note that if Evaluators were in the *No PR* condition then their final rating is their only rating. We use the term 'final' only to differentiate it sequentially from those Evaluators that also make a preliminary rating in the *PR* condition.

Employees, ensuring that statements did not contain a material misstatement of any fact, and reducing potential noise associated with variations in custom statements. Although not part of our theoretical development, we included the *SR+* condition for two reasons. First, it is a natural question from a mundane realism perspective as although our focus is on the influence of employees' quantitative self-ratings, as discussed in Section 2.1 concerning our survey of employees in practice, we find that in practice, employees often also provide both a self-rating and a justification for it. Second, research on persuasion suggests that providing a reason, even a weak one, when making a request can increase others' willingness to acquiesce to the request (Langer et al., 1978; Cialdini, 1993). Thus, the *SR+* condition allows us to test the robustness of our underlying theory with respect to the influence of self-ratings by investigating whether the effects we predict depend on the presence of a justification.

Our *Gender Identification* variable is a measured variable. Specifically, after providing their final rating of the Employee, Evaluators were asked a series of demographic questions, including a request to specify whether they identified primarily as male or female.⁶ For completing the study, Evaluators received fixed compensation of \$2 and bonus pay that varied based on the performance points generated by their assigned Employee and their final rating of the Employee. We included this bonus feature in order to establish a clear economic benchmark for Evaluators' rating behavior. Doing so allows us to better isolate the underlying behavioral effects that we predict. Specifically, Evaluators were paid an additional \$0.05 for each

⁶ We use a binary classification of gender for two reasons. First, the background theory and prior empirical research focus on male and female differences and thus we have limited grounds for predicting the behavior of individuals with other gender identities. Second, given that most individuals in the population identify as either male or female, it would be difficult to obtain a sufficient sample of individuals of other gender identifications to power meaningful statistical tests. Thus, rather than excluding participants on the basis of gender or complicating the subsequent analysis by adding multiple categories for gender, we simply asked participants “[t]o which gender do you most identify?” with “male” and “female” as the options. We would expect any gender identification errors to be randomized between males and females and across conditions.

performance point the Employee generated, consistent with the idea that in practice, better performance by Employees often also leads to better pay for the managers who evaluate them. However, Evaluators also had their compensation reduced as their employee rating score increased. Specifically, their compensation decreased by \$0.02 for each point on the evaluation scale, such that their compensation was decreased by \$0 if they rated the Employee a 1 and by \$0.20 if they rated the employee a 10. Thus, in all conditions, Evaluators have a financial disincentive to giving Employees high ratings. This disincentive provides three benefits. First, it provides a stronger test of our theory by biasing against leniency and in favor of strictness in rating performance. Second, although not a focus of our study, it establishes a clear economic benchmark for what constitutes purely self-interested Evaluator behavior. Third, it helps extend the applicability of our study to settings in which firms institute disincentives to deter leniency.

Evaluator participants

We obtained a performance evaluation from 863 Evaluators. Panel A of Table 1 reports their demographic characteristics. Evaluators report a mean age of 38.6 years, work experience of 16.9 years, and 54.1 percent identify as male. The median educational attainment reported was a four-year college degree, and the median reported income range is \$35,000 to \$74,999. Fifty-five percent of the participants report having some managerial experience. While our participants' demographics suggest they are likely very familiar with having their own performance evaluated by others, participants also report good familiarity with conducting performance evaluations, as 46.7 percent indicate that they have provided an evaluation of others' performance. Of these participants, 45 percent report conducting 1 to 5 evaluations

annually, 21.3 percent report conducting 6 to 10 evaluations annually, and the remaining 33.7 percent report conducting more than 10 evaluations annually.⁷

IV. RESULTS

Dependent and independent variables

Our primary dependent variable is *Rating*, which is the final rating of the Employee's performance by an Evaluator. A higher *Rating* implies Employee performance was evaluated more favorably than a lower *Rating*. We have three primary independent variables. *Gender* indicates whether the Evaluator identifies primarily as female (=0) or male (=1). *PR* indicates whether the Evaluator provided (=1) or did not provide (=0) a preliminary rating of the Employee. *SR* signifies whether the Evaluator did not receive the Employee's self-rating prior to providing a final rating of the Employee (=0, i.e., the No SR condition), did receive it (=1, i.e., the *SR* condition), or received it along with a message justifying the Employee's self-rating (=2, i.e., the *SR+* condition).

Descriptive statistics

Consistent with our *a priori* expectations, Employees ($M = 8.4$, untabulated) rate themselves more leniently than Evaluators ($M = 6.85$). Thus, Employees' self-ratings could plausibly upwardly anchor Evaluators' *Ratings*. Panel B of Table 1 presents additional descriptive statistics. Visual inspection of the mean *Rating* provided to Employees by Evaluators for each condition, as well as the Evaluator's self-identified *Gender*, yields a couple of observable patterns. Consistent with our predictions, mean *Ratings* are lower for males than for females (H1). Notably, this pattern appears to be quite persistent, as males' ratings are consistently lower than females' ratings across all six conditions. Also, *Ratings* are higher when Evaluators receive

⁷ An untabulated MANOVA indicates no significant differences occur in these demographic characteristics across the six conditions (Wilks' lambda = 0.8958, F = 0.79, p = 0.873)

a self-rating than when they do not, irrespective of *Gender* (H2). Also of note, the mean *Rating* appears similar regardless of whether the Evaluator provided a preliminary rating (PR) or did not do so (*No PR*).

Tests of hypotheses⁸

Our H1 prediction is that male evaluators will provide lower employee performance ratings than female evaluators. To test H1, we use all 863 independent Evaluator observations and conduct a t-test to compare the mean evaluation provided by males ($M = 6.58$) to that provided by females ($M = 7.17$). As shown in Panel A of Table 2, we find that males provide significantly lower ratings than females ($t = 3.87$, $p < 0.001$). Thus, the test results are consistent with H1: male evaluators were less lenient than female evaluators.⁹

H2 predicts that Evaluators will provide higher ratings when they receive an employee's self-rating prior to making their evaluation decision than when they do not receive it. Recall that we have two operationalizations of self-rating; one in which Evaluators received only the employee's numerical self-rating (*SR* condition) and a version in which they received both the numerical rating and a qualitative justification for the self-rating (*SR+* condition). Thus, prior to testing H2, we first analyze whether the mean evaluation in the *SR+* condition differs from that in the *SR* condition. Although the mean *Rating* in the *SR+* condition is directionally higher at

⁸ Reported p-values for the tests of the hypotheses are one-tailed or one-tailed equivalents (in the case of F tests) unless stated otherwise.

⁹ As discussed previously, evaluators practiced the same task that employees performed (also for 60 seconds). This allowed them to gain familiarity with the task requirements and potentially develop their own performance expectations. In untabulated analysis we find that during the Evaluator practice round male evaluators tended to complete more sliders than female evaluators ($M = 18.2$ vs. $M = 16.7$, $t = 4.5$, p -value < 0.001 , two-tailed). This could have led male evaluators to develop higher performance expectations than female evaluators and thus led them to provide relatively lower performance ratings for a given level of performance. To test whether this accounts for the effect of *Gender* we document in H1, we re-run H1's test as a regression with *Gender* as the independent variable and Evaluators' practice round task output as a covariate. Untabulated analysis indicates that while higher practice round output by evaluators leads to lower ratings ($t = 6.5$, $p < 0.001$, two-tailed), presumably due to having higher performance expectations, the effect of *Gender* remains statistically significant ($t = 2.92$, $p < 0.01$, one-tailed). Thus, we retain support for H1.

7.11 versus 6.92 in the *SR* condition, an untabulated t-test indicates that there is no significant difference ($t = 1.09$, $p = 0.13$, one-tailed). Essentially, evaluators' ratings of employees did not vary depending on whether employees provided a justification alongside their self-ratings. As such, to provide a simpler analysis, we pool the observations from the *SR* ($n = 288$) and *SR+* ($n = 287$) conditions into a single pool, which we hereafter refer to as *SR-all* ($n = 575$). Panel C of Table 1 reports the condensed summary data of the mean *Rating* by *Gender* and *SR-all* conditions using this pooled data. To test H2, we use a t-test to compare the mean *Rating* in the *SR-all* sample ($M = 7.02$) with that in the *No SR* condition ($M = 6.52$). As shown in Panel B of Table 2, we find that providing a self-rating led to significantly higher ratings by Evaluators ($t = 3.08$, $p = 0.001$). Thus, consistent with H2, we find that evaluators provide higher ratings when they receive an employee's self-rating prior to making their evaluation decision than when they do not receive it.

H3 predicts that the positive effect of receiving an employee's self-rating on evaluator ratings will be stronger when evaluators are male versus female. That is, H3 predicts an interaction between *Gender* and *SR*. To test H3, we conduct an ANOVA with *Rating* as the dependent variable and *SR-all*, *Gender*, and an interaction term *SR-all* \times *Gender* as the independent variables. As reported for Model 1 in Panel C of Table 2, we find that there is a marginally significant interaction ($F = 2.09$, one-tailed equivalent p -value = 0.074), consistent with H3. Further, as shown in Model 2, we find that when we include a control variable for the performance points generated by the Employee (as, *ceteris paribus*, employees with higher performance are likely to receive higher evaluations), this interaction term strengthens to conventional levels of significance ($F = 3.03$, one-tailed equivalent p -value = 0.041). To help illustrate the nature of the interaction, Figure 2 plots the mean *Rating* by *Gender* and *SR-all*. The

analysis for H3 also provides a secondary test of H1 and H2, as *Gender* and *SR-all* serve as main effects in the ANOVA. Consistent with our primary test of H1, Model 1 indicates that *Gender* is significant ($F = 17.08$, $p < 0.001$), while, consistent with H2, *SR-all* is also significant ($F = 8.67$, $p = 0.003$). We expand our test of H3 by examining the simple effect of self-rating for males and females. As reported in Panel D of Table 2, we find that receiving a self-rating does not significantly affect evaluations given by females ($F = 1.03$, $p = 0.310$) but does significantly affect those given by males ($F = 10.55$, $p = 0.001$). This provides additional evidence consistent with H3 concerning the differential effects of employee self-ratings on male and female evaluators.

Finally, additional analysis of the simple effect of *Gender* indicates that males still provide lower ratings than females irrespective of whether they receive an employee's self-rating ($F = 5.41$, p -value = 0.020) or do not receive it ($F = 11.67$, $p < 0.001$). Thus, while receiving employees' self-ratings has stronger upward effects on the ratings provided by males than females, males nonetheless continue to provide lower ratings than females. In sum, we find support for our three hypotheses. Males provide lower ratings than females. Providing employees' self-ratings to evaluators leads to higher overall ratings, and this effect is more pronounced among male evaluators than female evaluators. Indeed, as the simple effects analysis demonstrates, it is only present among males and not among females.

Additional analysis – Preliminary ratings

Having conducted our primary tests, we now investigate whether our results are robust to evaluators providing an explicit preliminary rating of an employee's performance before receiving the employee's self-rating, and whether this additional step in the evaluation process would mitigate any effects of providing such ratings to evaluators. It is important to examine the

robustness of our predicted behavior by examining whether a potential intervention exists that could mitigate the leniency-inducing effect of employees' self-assessments while still affording employees a voice in their evaluation. We focus on preliminary ratings by the evaluator, in particular because they also fit within our theoretical framework of anchoring and adjustment, now considering the potential influence of an additional anchor point. Specifically, as part of the evaluation process, evaluators can form *preliminary* ratings about an employee's performance based on their assessment of the objective performance measures and *before* receiving any subjective self-assessments from the employee. Given that we predict evaluators anchor their judgments on employees' self-assessments in the absence of forming any preliminary ratings, this allows us to examine whether having evaluators establish an initial anchor point might alter their reliance on employees' self-assessments.

Hogarth and Einhorn's (1992) belief-adjustment model outlines the underlying uncertainty regarding how individuals might respond to multiple potential anchors in their decision making process. Specifically, the process individuals engage in to weigh multiple relevant pieces of information can vary as some evaluators might consider both evaluations but anchor primarily on their initial preliminary rating (i.e., what prior research identifies as a primacy effect) while adjusting to consider the employees' self-assessment. Alternatively, other evaluators could focus primarily (or even exclusively) on the most recent anchor (i.e., what prior research identifies as a recency effect), thereby largely replacing their initial assessment of the employee with the employee's self-assessment. Finally, some evaluators might take a more balanced approach, weighing both evaluations more equally when determining their ultimate evaluation.

We performed several analyses to examine the impact, if any, that having evaluators provide an initial rating has on our primary results. To do so, we use the sample of cases in which evaluators receive the employee's self-rating, which captures the cases in which the employee's self-rating could serve as a possible anchor for the evaluator (i.e., we use the observations for *SR-all* as defined above). Using this sample, we first conduct a t-test to determine if the overall mean *Rating* for evaluators differs by *PR* condition (i.e., whether the evaluator forms a preliminary rating of the Employee before receiving the employee's self-rating). We find no evidence that the mean *Rating* given by evaluators between the *PR* ($M = 7.02$) and *No PR* ($M = 7.01$) conditions differs significantly ($t = 0.10$, p -value = 0.924, two-tailed, untabulated). In the context of Hogarth and Einhorn's (1992) belief-adjustment model, evaluators' apparent primary reliance on the employees' self-rating over their own preliminary rating is most consistent with a recency-type approach to weighting multiple potential anchor points.

To gain insights regarding whether evaluators' preliminary ratings affect the takeaways from our primary analysis, we retest each of our three hypotheses using only the subset of evaluators in the *PR* condition. Untabulated results suggest that our primary results regarding the influence of employees' self-ratings and the role of gender are robust even in settings where evaluators hold explicit beliefs about an employee's performance prior to receiving performance evaluation input from the employee. Specifically, regarding H1, untabulated analysis indicates male evaluators remain less lenient than female evaluators ($t = 2.55$, $p < 0.01$, one-tailed) whereas, regarding H2, evaluators continue to provide higher ratings when they receive an employee's self-rating prior to making their evaluation decision than when they do not receive it ($t = 2.31$, $p < 0.01$, one-tailed). Finally, regarding our H3 interaction prediction that the effect of an employee's self-rating on evaluator ratings will be stronger when evaluators are male versus

female, both Models 1 and 2 presented above remain marginally significant (Model 1: $F = 1.80$, one-tailed p-equivalent value = 0.094; Model 2: $F = 2.70$, one-tailed p-value equivalent = 0.05) when considering only the subset of evaluators that were required to provide a preliminary rating.

V. CONCLUSION

Leniency in performance ratings is a well-documented phenomenon of subjective performance evaluations. Leniency imposes multiple costs on firms, including making it difficult to distinguish between high-performers and low-performers for the purpose of providing meaningful feedback, making informed promotion decisions, motivating future employee effort, and allocating resources effectively for employee training. Given these costs, it is important to identify and better understand the underlying factors that can cause evaluators to provide higher performance ratings.

Our study highlights that a common feature of performance rating processes, namely having employees self-rate their own performance and incorporating this into the information used to evaluate employees, is an underlying driver of ratings leniency. We document this using a rather understudied yet practically relevant setting in which managers make evaluation decisions in the face of performance information noise. Allowing employees to provide quantitative numerical self-ratings can convey several benefits to firms, particularly in settings like ours where there is performance-measure noise. We predict and show that it can also foster higher ratings due to the propensity of flattering self-ratings to act as anchors on evaluators' own judgments. Further, we develop theory explaining why males and females will handle uncertainty in performance information differently, and we also provide corresponding evidence that the effect of employee self-ratings on evaluator leniency is more pronounced among male

evaluators than female evaluators. Male evaluators, while generally stricter in their performance ratings, provided a larger increase in ratings when presented with employee self-ratings than their female counterparts. Indeed, we find no empirical evidence that receiving employees' self-ratings induces female evaluators to provide significantly higher ratings.

Prior research has examined numerous relevant and important factors that can potentially influence evaluators' leniency in their performance ratings of employees. Broadly speaking, research has studied (i) the psychological and personality forces that foster it (e.g., Tziner et al. 2005; Cheng et al. 2017; Jawahar 2001), (ii) the economic incentives and tradeoffs associated with it (e.g., Wayne and Liden 1995; Bol 2011), (iii) demographic characteristics of those being evaluated that can affect it (e.g., Heilman 1983; Greenhaus et al. 1990; Ragins and Cornwell 2001; Lyness and Heilman 2006; Roberson and Kulik 2007; De Pater et al. 2010; King et al. 2010), and (iv) contextual features of the evaluation process itself that may promote it (e.g., Farh et al. 1991; Tziner et al. 2005; Yun et al. 2005; Joshi and Roh 2009; Tan 2019; Gong et al. 2021). As discussed below, our study contributes to and extends the broad research stream on leniency in relevant, consequential, and interesting ways.

First, regarding gender (and other demographic characteristics), the majority of the focus has been on employees' gender, rather than evaluators' gender. Thus, to our knowledge, our study is (one of) the first to investigate the relevance of gender role theory from the perspective of evaluators' performance ratings and provide empirical evidence with respect to gender and employee rating leniency. Second, our study identifies a potential unintended cost, in the form of ratings leniency, associated with the common contextual feature of having employees provide self-ratings to their evaluator. Our focus and insights are novel as although our survey data indicate numerical self-ratings are common in firms' evaluation processes, they are typically

framed and studied in terms of what they can add to the process. Third, and most importantly, we provide insights on why and how these two factors interact to ultimately affect evaluator ratings.

Not only are insights into this interaction important for theory-building, but they also have practical implications. Our study highlights the need for firms to recognize that although incorporating 'self-ratings' into their ratings process can convey benefits to the firm and heighten the perceived legitimacy of the ratings process, it could also impose costs associated with more lenient performance ratings, including decreased ratings legitimacy. Moreover, our results suggest that even if evaluators already have pre-existing opinions about employees' performance, they can still be influenced by employees' self-ratings. Our study also suggests that this problem is likely to be more prevalent in firms in which mostly males rate performance, which has tended to be the case for most firms historically and continues to be the case for a majority of firms today (Eagly and Carli 2007; Grossman et al. 2019; Catalyst 2023). Ultimately, our study highlights the complex interplay between evaluator gender, employee self-ratings, and leniency in performance ratings when there is performance measure noise. In doing so, we can help firms better identify and understand the relevant tradeoffs in this common setting. A better understanding of gender-based differences in performance ratings and how these differences interact with a prominent feature of the ratings process can help firms design more effective rating systems and incorporate mechanisms to mitigate potential biases.

Future research could investigate multiple extensions of this study. First, we made several experimental design decisions to improve control and better test our underlying theory. For example, we were silent about the gender identification of the employee being rated. However, prior research finds that employee gender can matter in employee ratings, as well as hiring and promotion decisions (e.g., Ridgeway 2001; Eagly and Karau 2002; Bowles and Gelfand 2010;

Heilman 2012; Rice and Barth 2016; Chandler, 2018; Fanning et al. 2021). Therefore, a relevant and interesting extension would be to expand the gender dynamics of our study by investigating how, if at all, the results we document would change if evaluators are cognizant of their employees' gender.

Second, our setting purposely did not provide evaluators with any explicit corporate policy or guidelines (e.g., forced ranking requirements, historical ranges of past evaluations, etc.) so as not to artificially constrain evaluators' decision making in our setting. We also excluded any direct interaction between evaluators and evaluatees pre- or post-evaluation, while framing evaluators' decisions as hypothetical, in that they ultimately do not affect the employee being evaluated. In doing so, we can rule out our results being driven by factors such as differential adherence to expectations, evaluators' fear of confrontation with their employees regarding the evaluation, evaluators' concerns about how their decisions might affect employees' pay, and reputation-related effects from either party based on past or ongoing interactions. While these design decisions abstract our setting from one with more mundane realism, controlling for these potential influences allows us to better test our underlying theory. We acknowledge that, to the extent that any of these factors influence the interaction between self-ratings and evaluator gender that we investigate and document, they would provide additional worthwhile insights. However, given that these are beyond the scope of our study and that such inferences are not immediately clear *ex ante*, we consider them relevant areas for future research.

Third, having documented that gender and employee self-ratings can interact to foster higher evaluator ratings, it would be interesting to know whether cost-effective interventions exist that can mitigate these effects while still allowing employees to provide input during their evaluations. For instance, training programs aimed at raising awareness of gender biases can help

evaluators recognize and counteract their own biases (Bohnet 2016). Such programs might help firms better understand and manage evaluators' anchoring effects or even implement structured guidelines for integrating self-ratings into final evaluations. Additionally, while incorporating an expanded self-rating process, multiple evaluators via a calibration-type committee, and/or 360-degree feedback systems can provide a more comprehensive and balanced view of employee performance (Brett and Atwater 2001), perhaps they can also help reduce the impact of the anchoring-based effects we document. Moreover, examining the role of other demographic variables and contextual factors in moderating these effects could provide a more comprehensive understanding of how to improve performance rating processes.

Fourth, future research could investigate whether our theoretical predictions pertaining to anchoring and adjustment, as well as our related findings, hold in settings where evaluators' evaluations of employees, as well as employees' self-assessments, are more qualitative in nature (i.e., not primarily quantified via a ratings scale). Finally, although we rely on an abstract task in our experiment for theory-testing purposes, experience in providing domain-specific performance ratings could influence our results. Though beyond the scope of this study, future research could investigate how more extensive experience rating performance, as well as the domain-specific characteristics of tasks, could affect our results.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used ChatGPT 5 in order to assist with identifying existing research studies that might have relevance for the current study's motivation, theory development, and / or incremental contribution. Upon receiving the AI-generated output, the authors reviewed the identified studies to determine potential relevance. The authors take full responsibility for the content of the publication.

Appendix

Survey Questions and Responses

Have you received a performance evaluation as part of your employment in the last three years? (105 participants)

Yes: 88 participants (83.8%)
No: 17 participants (16.2%)

What information, if any, were you asked to provide to your evaluator as part of your performance evaluation? (88 participants)

54 participants (61.4%) were asked to provide a self-assessment of their performance.
55 participants (63.3%) were asked to provide a summary of their work accomplishments during the period.
2 participants (2.3%) indicated they were asked to provide other types of information (e.g., goals for the next year, feedback on the evaluation).

Were you asked to self-rate your performance on a numerical ratings scale (e.g., a 1 to 5 scale)? (54 Participants)

Yes: 41 participants (75.9%)
No: 13 participants (24.1%)

Was the scale you used the same scale that your evaluator ultimately used to evaluate your performance? (41 participants)

Yes: 40 participants (97.6%)
No: 1 participant (2.4%)

Were you asked to justify your self-assessment (i.e., explain why you think you deserved that rating)? (54 participants)

Yes: 47 participants (87.0%)
No: 7 participants (13.0%)

Were you asked to provide materials to support your self-assessment? (54 participants)

Yes: 16 participants (29.6%)
No: 38 participants (70.4%)

References

- Atwater, L. E., and F. J. Yammarino 1997. Self-other rating agreement: A review and model. *Research in Personnel and Human Resources Management* 15: 121–174.
- Baker, G. P., R. Gibbons, and K. J. Murphy. 1994. Subjective performance measures in optimal incentive contracts. *Quarterly Journal of Economics* 109 (4): 1125–56.
- Bem, S. L. 1981. Gender schema theory: A cognitive account of sex typing. *Psychological Review* 88 (4): 354–364.
- Bohnet, I. 2016. *What works: Gender equality by design*. Harvard University Press.
- Bol, J. C. 2008. Subjectivity in compensation contracting. *Journal of Accounting Literature* 27: 1–24.
- Bol, J. C. 2011. The determinants and performance effects of managers' performance evaluation biases. *The Accounting Review* 86 (5): 1549–1575.
- Bol, J. C., and S. D. Smith. 2011. Spillover effects in subjective performance evaluation: Bias and the asymmetric influence of controllability. *The Accounting Review* 86 (4): 1213–1230.
- Bol, J. C., S. Kramer, and V. S. Maas. 2016. How control system design affects performance evaluation compression: The role of information accuracy and outcome transparency. *The Accounting Review* 51: 64–73.
- Bol, J. C., A. B. De Aguiar, and J. B. Lill. 2025. Calibration in the Performance Evaluation Process. *Human Resource Management* 64 (4): 1141–1159.
- Bonner, S. E., and G. B. Sprinkle. 2002. The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Accounting, Organizations and Society* 27 (4–5): 303–45.
- Bosak, J. and S. Sczesny. 2011. Gender bias in leadership selection? Evidence from a hiring simulation study. *Sex Roles* 65, 234–242.
- Bowles, H. R., and M. J. Gelfand. 2010. Status and the evaluation of workplace deviance. *Psychological Science* 21 (1): 49–54.
- Brett, J. F., and L. E. Atwater. 2001. 360-Degree feedback: Accuracy, reactions, and perceptions of usefulness. *Journal of Applied Psychology* 86 (5): 930–942.
- Bussey, K. and A. Bandura. 1999. Social cognitive theory of gender development and differentiation. *Psychological Review* 106 (4): 676–713.
- Catalyst 2023. *Pyramid: Women in the United States workforce*. February 07, 2023. Catalyst.org
- Chan, E. 2017. Promotion, relative performance information, and the peter principle. *The Accounting Review* 93 (3): 83–103.
- Chandler, V. 2018. Do evaluators prefer candidates of their own gender? *Canadian Public Policy* 44 (4): 289–302.
- Cheng, K. H. C., C. H. Hui, and W. F. Cascio. 2017. Leniency bias in performance ratings: The big-five correlates. *Frontiers in Psychology* 8 (April), 1–10.
- Cialdini, R. B. 1993. *Influence: Science and practice* (3rd ed.). HarperCollins College Publishers.
- Cristofaro, M., and P. L. Giardino. 2020. Core self-evaluations, Self-leadership, and the self-serving bias in managerial decision making. *Administrative Sciences* 10 (3): 64.
- De Pater, I. E., A. E. Van Vianen, and M. N. Bechtoldt. 2010. Gender differences in job challenge: A matter of task allocation. *Gender, Work and Organization* 17 (4): 433–453.
- Demere, B. W., K. L. Sedatole, and A. Woods. 2019. The role of calibration committees in subjective performance evaluation systems. *Management Science* 65 (4): 1562–1585.

- Dunning, D., C. Heath, and J. M. Suls. 2004. Flawed self-assessment implications for health, education, and the workplace. *Psychological Science in the Public Interest* 5(3): 69–106.
- Dunning, D., K. Johnson, J. Ehrlinger, and J. Kruger. 2003. Why people fail to recognize their own incompetence. *Current Directions in Psychological Science* 12 (3): 83-87.
- Eagly, A. H., and L. L. Carli. 2007. Women and the labyrinth of leadership. *Harvard Business Review* 85 (9): 62-71.
- Eagly, A. H., and S. J. Karau. 2002. Role congruity theory of prejudice toward female leaders. *Psychological Review* 109 (3): 573-598.
- Eagly, A. H., and W. Wood. 2012. Social role theory. In P. A. M. Van Lange, A. W. Kruglanski, and E. T. Higgins (Eds.), *Handbook of Theories of Social Psychology* (pp. 458-476). Sage Publications.
- Epley, N., and T. Gilovich. 2001. Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science* 12 (5): 391-396.
- Epley, N., and T. Gilovich. 2006. The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science* 17 (4): 311-318.
- Erdogan, B., M. L. Kraimer, and R. C. Liden. 2001. Procedural justice as a two-dimensional construct: An examination in the performance appraisal context. *Journal of Applied Behavioral Science* 37 (2): 205-222.
- Fanning, K., J. Williams, and M. Williamson. 2021. Group recruiting events and gender stereotypes in employee selection. *Contemporary Accounting Research* 38 (4): 2496-2520.
- Farh, J. L., and G. H. Dobbins. 1989. Effects of self-Esteem on leniency bias in self-reports of performance: A structural equation model analysis. *Personnel Psychology* 42 (4): 835–850.
- Farh, J. L., A. A. Cannella, and A. G. Bedeian. 1991. Peer ratings: The impact of purpose on rating quality and user acceptance. *Group and Organization Studies* 16 (4): 367-386.
- Feingold, A. 1994. Gender differences in personality: A meta-analysis. *Psychological Bulletin* 116 (3): 429–456.
- Fletcher, C. 1999. The implications of research on gender differences in self-assessment and 360 degree feedback. *Human Resource Management Journal* 9 (1): 39–46.
- Folger, R., M. A. Konovsky, and R. Cropanzano. 1992. Due process metaphor for performance appraisal. *Research in Organizational Behavior* 14: 129-177.
- Fulk, J., A. P. Brief, and S. H. Barr. 1985. Trust-in-supervisor and perceived fairness and accuracy of performance evaluations. *Journal of Business Research* 13: 301–313.
- Fuhrmans, V. (2024, September 17). A decade after 'lean in,' progress for women isn't trickling down. *The Wall Street Journal*. <https://www.wsj.com/lifestyle/careers/a-decade-after-lean-in-progress-for-women-isnt-trickling-down-f0e34074?mod=djem10point>
- Furnham, A., and H. C. Boo. 2011. A literature review of the anchoring effect. *The Journal of Socio-Economics* 40 (1): 35-42.
- Gibbs, M., K. A. Merchant, W. A. Van der Stede, and M. E. Vargus. 2004. Determinants and effects of subjectivity in incentives. *The Accounting Review* 79 (2): 409–36.
- Gill, D., and V. Prowse. 2015. A novel computerized effort task based on sliders. Working paper, University of Oxford and Cornell University.
- Golman, R. and S. Bhatia. 2012. Performance evaluation inflation and compression. *Accounting, Organizations, and Society* 37 (8): 534–543.

- Gong, N., W. F. Boh, A. Wu, and T. Kuo. 2021. Leniency bias in subjective performance evaluation: Contextual uncertainty and prior employee performance, *Emerging Markets Finance and Trade* 57 (8): 2176-2190.
- Greenhaus, J. H., S. Parasuraman, and W. M. Wormley. 1990. Effects of race on organizational experiences, job performance evaluations, and career outcomes. *Academy of Management Journal* 33 (1): 64–86.
- Grossman, P. J., C. Eckel, M. Komai, and W. Zhan. 2019. It pays to be a man: Rewards for leaders in a coordination game. *Journal of Economic Behavior and Organization* 161: 197-215.
- Harris, M. M., and J. Schaubroeck. 1988. A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology* 41 (1): 43-62.
- Heilman, M. E. 1983. Sex Bias in Work Settings: The lack of fit model. *Research in Organizational Behavior* 5: 269-298.
- Heilman, M. E. 2012. Gender stereotypes and workplace bias. *Research in Organizational Behavior* 32: 113-135.
- Hoffman, M. L. 1977. Sex differences in empathy and related behaviors. *Psychological Bulletin* 84 (4): 712–722.
- Hogarth, R. M., and H. J. Einhorn. 1992. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology* 24: 1-55.
- Hoogervorst, N., D. De Cremer, and M. van Dijke. 2013. "Fairness and the power of giving voice to employees." *Journal of Managerial Psychology* 28 (5): 535-554.
- Jawahar, I. M. 2001. Attitudes, self-monitoring, and appraisal behaviors. *Journal of Applied Psychology* 86 (5): 875–883.
- Joshi, A., and H. Roh. 2009. The role of context in work team diversity research: A meta-analytic review. *Academy of Management Journal* 52 (3): 599–627.
- Judge, T. A., and J. E. Bono. 2001. Relationship of core self-evaluations traits: Self-esteem, generalized self-efficacy, locus of control, and emotional stability with job satisfaction and job performance: A meta-analysis. *The Journal of Applied Psychology* 86: 80–92.
- Karagozoglu, E., and A. Riedl. 2015. Performance information, production uncertainty, and subjective entitlements in bargaining. *Management Science* 61 (11): 2549-2824.
- Kelly, K., T. Vance, and R. A. Webb. 2015. The interactive effects of *ex post* goal adjustment and goal difficulty on performance. *Journal of Management Accounting Research* 27 (1): 1–25.
- King, E. B., M. R. Hebl, J. M. George, and S. F. Matusik. 2010. Understanding tokenism: Antecedents and consequences of a psychological climate of gender inequity. *Journal of Management* 36 (2): 482-510.
- Kraiger, K., and J. K. Ford. 1985. A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology* 70 (1): 56-65.
- Kromrei, H. 2015. Enhancing the annual performance appraisal process: Reducing biases and engaging employees through self-assessment. *Performance Improvement Quarterly* 28 (2): 53–64.
- Kruger, J., and D. Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77 (6): 1121-1134.

- Langer, E., A. Blank, and B. Chanowitz. 1978. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology* 36 (6): 635-642.
- Libby, R., R. Bloomfield, and M. W. Nelson. 2002. Experimental research in financial accounting. *Accounting, Organizations, and Society* 27 (8): 775–810.
- Lippa, R. A. 2005. *Gender, nature, and nurture*. Routledge.
- Luft, J., M. D. Shields, and T. F. Thomas. 2016. Additional information in accounting reports: Effects of management decisions and subjective performance evaluations under causal ambiguity. *Contemporary Accounting Research* 33 (2): 526–550.
- Lyness, K. S., and M. E. Heilman. 2006. When fit is fundamental: Performance evaluations and promotions of upper-level female and male managers. *Journal of Applied Psychology* 91 (4): 777-785.
- Meyer, H. H. 1991. A solution to the performance appraisal feedback enigma. *The Executive* 5 (1): 68–76.
- Moers, F. 2005. Discretion and bias in performance evaluations: The impact of diversity and subjectivity. *Accounting, Organizations, and Society* 30 (1): 67–80.
- Morrison, E. W., and F. J. Milliken. 2000. Organizational silence: A barrier to change and development in a pluralistic world. *Academy of Management Review* 25 (4): 706–725.
- Ragins, B. R., and J. M. Cornwell. 2001. Pink Triangles: Antecedents and consequences of perceived workplace discrimination against gay and lesbian employees. *Journal of Applied Psychology* 86 (6): 1244-1261.
- Rice, L. and J. M. Barth. 2016. Hiring decisions: The effect of evaluator gender and gender stereotype characteristics on the evaluation of job applicants. *Gender Issues* 33: 1–21.
- Ridgeway, C. L. 2001. Gender, status, and leadership. *Journal of Social Issues*, 57(4), 637–655.
- Roberson, Q. M., and C. T. Kulik. 2007. Stereotype threat at work. *Academy of Management Perspectives* 21 (2): 24-40.
- Rudman, L. A. 1998. Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology* 74 (3): 629-645.
- Rudman, L. A., and P. Glick. 2001. Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues* 57 (4): 743-762.
- Sedikides, C., and A. P. Gregg. 2008. Self-enhancement: Theory and research. In O. P. John, R. W. Robins, and L. A. Pervin (Eds.), *Handbook of Personality: Theory and Research* (3rd ed., pp. 110-138). Guilford Press.
- Tan, Boon-Seng 2019. In search of the link between organizational culture and performance: A review from the conclusion validity perspective. *Leadership and Organization Development Journal* 40 (3): 356-368.
- Tversky, A., and D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185 (4157): 1124-1131.
- Tziner, A., K. R. Murphy, and J. N. Cleveland. 2005. Contextual and rater factors affecting rating behavior. *Group and Organization Management* 30 (1): 89-98.
- Wayne, S. J., and R. C. Liden. 1995. Effects of impression management on performance ratings: A longitudinal study. *Academy of Management Journal* 38 (1): 232-260.
- Wick, S. 2021. Subjectivity in performance evaluations: A review of the literature. *Accounting Perspectives* 20 (4): 653-685.

Yun, G. J., L. M. Donahue, N. M. Dudley, and L. A. McFarland. (2005). Rater personality, rating format, and social context: Implications for performance appraisal ratings. *International Journal of Selection and Assessment* 13 (2): 97–107.

Figure 1 – Illustration of Employee task

Slide to 30:

0 10 20 30 40 50 60 70 80 90 100

Slide to 30



Slide to 70

0 10 20 30 40 50 60 70 80 90 100

Slide to 70



Slide to 10

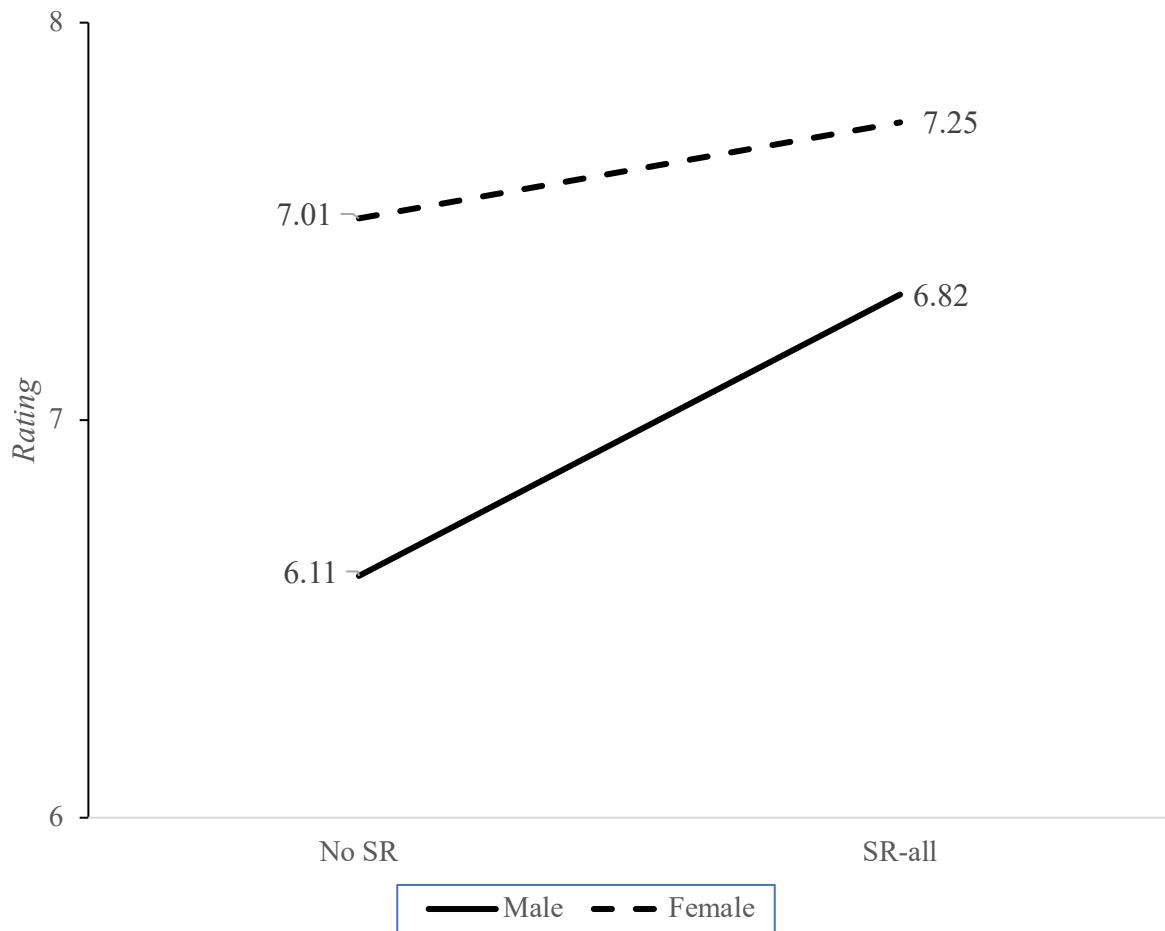
0 10 20 30 40 50 60 70 80 90 100

Slide to 10



Notes on Figure 1. Employees were tasked with using the mouse on their computer to click on the red dot located on the left side of each line and then dragging the circle to the indicated line position. Each dot correctly positioned generated two performance points for the Employee.

Figure 2 – Effect of self-ratings on *Rating* by Gender



Notes on Figure 2. *Rating* is the performance evaluation rating that Evaluators gave to Employees. Evaluators used an 11-point ratings scale bounded by 1 (Poor performance) and 10 (Outstanding performance). *SR-all* (*No SR*) consists of observations of Evaluators who were (not) provided with the Employee's performance self-rating prior to the Evaluator making their rating. Thus, *SR-all* combines the *SR* and *SR+* conditions.

Table 1 – Descriptive statistics**Panel A. Demographics of Evaluators**

n = 863

Variable	Mean	Median	Frequency
Age (years)	38.6		
Work experience (years)	16.9		
Proportion identifying as Male			54.10%
Educational attainment		4-year degree	
Income range		\$35,000 - \$74,999	
Proportion reporting managerial experience			55%
Proportion who have evaluated others*			46.70%
*Number of evaluations provided annually			
1 to 5			45%
6 to 10			21.30%
More than 10			33.70%

Panel B. Descriptive statistics about Ratings

		Gender of Evaluator								
		Male	(Std. Dev.)	n	Female	(Std. Dev.)	n	All	(Std. Dev.)	n
No PR condition	No SR	6.12	(2.53)	76	6.99	(2.26)	68	6.53	(2.44)	144
	SR	6.68	(2.35)	80	7.08	(2.25)	64	6.85	(2.31)	144
	SR+	6.85	(2.44)	72	7.48	(1.76)	71	7.16	(2.15)	143
Provide PR condition	No SR	6.10	(2.55)	81	7.03	(1.87)	63	6.51	(2.32)	144
	SR	6.78	(2.05)	78	7.21	(1.72)	66	6.98	(1.91)	144
	SR+	6.96	(2.17)	80	7.20	(2.50)	64	7.07	(2.32)	144
All conditions		6.58	(2.37)	467	7.17	(2.07)	396	6.85	(2.25)	863

Panel C. Mean Ratings of Evaluators, condensed

	Male	(Std. Dev.)	n	Female	(Std. Dev.)	n	Difference
No SR	6.11	(2.54)	157	7.01	(2.07)	131	-0.90
SR & SR+	6.82	(2.25)	310	7.25	(2.07)	265	-0.43
Difference	-0.71			-0.24			

Notes on Panels B and C: In the *PR* (*No PR*) condition, Evaluators did (did not) provide a preliminary evaluation of the Employee prior to providing a final evaluation. In the *No SR* (*SR*) condition, Evaluators did not (did) receive the Employee's self-rating prior to providing their final rating. In the *SR+* condition, Evaluators did receive the Employee's self-rating and a justification of that rating prior to providing their final rating.

Table 2 – Tests of hypotheses

Panel A. Test of H1 comparing mean *Ratings* by *Gender*

Gender	n	Mean	Standard Deviation
Female	396	7.17	2.07
Male	467	6.58	2.37
Difference		0.59	

t = 3.87

p-value < 0.001, one-tailed

Panel B. Test of H2 comparing mean Ratings by presence or absence of Employees' self-ratings.

Workers' self-ratings	n	Mean	Standard Deviation
Absent	288	6.52	2.38
Present	575	7.02	2.17
Difference		-0.50	

t = 3.08

p-value = 0.001, one-tailed

Table 2 continues on the next page.

Panel C. ANOVA test of H3, interaction of *Gender* and Employees' self-ratings

Source - Model 1	df	Mean Square	F-ratio	p-value
<i>Gender</i>	1	84.51	17.08	<0.001
<i>Self-rating</i>	1	42.91	8.67	0.003
<i>Gender</i> × <i>Self-rating</i>	1	10.36	2.09	0.074*
Error	859	4.95		

*reported p-value is one-tailed equivalent due to directional prediction

Source - Model 2	df	Mean Square	F-ratio	p-value
<i>Gender</i>	1	81.31	16.96	<0.001
<i>Self-rating</i>	1	43.03	8.97	0.003
<i>Gender</i> × <i>Self-rating</i>	1	14.55	3.03	0.041*
<i>Performance points</i>	17	12.49	2.60	<0.001
Error	842	4.80		

*reported p-value is one-tailed equivalent due to directional prediction

Notes on Panel C. *Gender* takes the value of 0 (1) if the Employee identifies as female (male). *Self-rating* takes the value of 0 (1) if the Evaluator did not (did) receive the Employee's self-rating prior to providing a final performance rating. For purposes of this analysis, the value of 1 combines the *SR* and *SR+* conditions as presented in Panel C of Table 1. *Gender*×*Self-rating* is an interaction variable between *Gender* and *Self-rating*.

Panel D. Simple effects related to H3

Effect of receiving Worker's self-rating by Evaluator's Gender

Gender	df	F	p-value
<i>Females</i>	1	1.03	0.310
<i>Males</i>	1	10.55	0.001
Joint	2	5.79	0.003
Denominator	859		

Effect of Evaluator's Gender by provision of Employee's self-rating

Self-rating	df	F	p-value
Not Provided	1	11.67	<0.001
Provided	1	5.41	0.020
Joint	2	8.54	<0.001
Denominator	859		