"Potential" and the Gender Promotion Gap*

Alan Benson
Univ. of Minnesota

Danielle Li
MIT & NBER

Kelly Shue
Yale & NBER

May 9, 2022

Please see here for latest version

Abstract

We show that widely-used subjective assessments of employee "potential" contribute to gender gaps in promotion and pay. Using data on 29,809 management-track employees from a large North American retail chain, we find that women receive substantially lower potential ratings despite receiving higher job performance ratings. Differences in potential ratings account for 30-50% of the gender promotion gap. Women's lower potential ratings do not appear to be based on accurate forecasts of future performance or attrition: women subsequently outperform male colleagues with the same potential ratings, both on average and on the margin of promotion, and women are less likely to exit the firm. Despite this, women's subsequent potential ratings remain low, suggesting that firms persistently underestimate the potential of their female employees.

JEL Classifications: M51 (Firm employment decisions; promotions); J71 (Discrimination); J31 (Wage level and structure; wage differentials); D84 (Expectations; speculations)

Keywords: Promotions, performance evaluations, glass ceiling, gender bias, leadership, compensation, wage differentials, role congruity theory, Peter Principle

*Alan Benson: bensona@umn.edu. Danielle Li: d_li@mit.edu. Kelly Shue: kelly.shue@yale.edu. We thank Yi Wang and Geng Li for excellent research assistance and the International Center for Finance at the Yale School of Management for their support. We thank Heather Sarsons for helpful discussion comments, and seminar audiences at University of Alberta, ASU, Bocconi, Columbia, CUHK, FOM Conference, INSEAD, Kellogg, Minnesota, MIT, NBER Personnel Economics, SIOE, Southern Methodist University, University of Maryland, University of Southern California, University of South Carolina, Warwick, and Yale for their feedback.

1 Introduction

When making promotion decisions, firms must form predictions about the future performance of each employee: if given the opportunity, would someone make a good manager? To guide this assessment, firms may use information about a worker's past job performance. Past performance, however, is an imperfect predictor of future performance both because people may change over time and because higher-level roles may require a different skillset (Peter and Hull, 1969; Benson, Li and Shue, 2019). As a result, many firms ask managers to make inferences about a worker's "potential." Yet because potential is not directly observed, these assessments can be highly subjective, leaving room for bias.

Prior studies highlight several reasons why subjective assessments may negatively impact the careers of women in particular. First, research on role congruity theory (Eagly and Karau, 2002) provides broad evidence that people find it difficult to imagine women as leaders because the qualities stereotypically associated with effective leaders—e.g. assertiveness, competitiveness, ambition, and an orientation toward execution—are also stereotypically associated with men.[1] Player et al. (2019), for instance, find that experimental subjects forecast greater leadership potential and future performance for male applicants whose resumes are otherwise identical to women. These stereotypes may also be reinforced by real-world differences in job and task assignments: women are much less frequently observed in managerial roles (Blau and Kahn, 2017) and are more likely to be asked to volunteer for service tasks that are not valued in the promotion process (Babcock et al., 2017). Second, subjective assessments may be heavily influenced by politicking and familiarity (Milgrom, 1988; Prendergast and Topel, 1993; Bertrand, Chugh and Mullainathan, 2005). This may generate gender-based disparities if women have less access to networking opportunities (e.g. Cullen and Perez-Truglia, 2020), benefit less from connections (e.g. Fang and Huang, 2017), or advocate less for their advancement (e.g. Babcock and Laschever, 2009; Biasi and Sarsons, 2020; Roussille, 2020).[2] Finally, self-interested managers may manipulate potential ratings to keep their best subordinates (Friebel and Raith, 2013). Haegele (2021) find that such "talent hoarding" leads to disproportionately lower promotion rates for women, possibly because female subordinates have a stronger distaste for confrontation with their managers.

[1] See also Bursztyn, Fujiwara and Pallais (2017); Koenig et al. (2011); Proudfoot, Kay and Koval (2015); Correll et al. (2020); and Kaplan, Klebanov and Sorensen (2012).

[2] Gender differences in these behaviors may also be driven by lower self-confidence (Sarsons and Xu, 2021) or lower aspirations (Azmat, Cuñat and Henry, 2020). We note that possible gender differences in confidence and aspirations are compatible with a view in which promotions are biased against women. Azmat, Cuñat and Henry (2020), for example, model aspirations as endogenously determined by workplace gender discrimination (see also Brands and Fernandez-Mateo (2017)). In our setting, women receive higher performance ratings than men, suggesting that any gender differences in aspirations do not translate into lower effort provision by women.

We show that subjective assessments of worker potential contribute to gender gaps in promotion and to an inefficient allocation of talent across roles. We study promotions among 29,809 management-track workers within a large North American retail chain. Our firm uses a popular talent assessment and succession planning tool known as a "Nine Box" grid, in which direct supervisors rate subordinates on two dimensions: their current performance and their future potential. Each dimension takes three values (1-low, 2-medium, and 3-high), creating a 3-by-3 matrix with nine cells. Whereas performance ratings are intentionally backward-looking and often based on demonstrable achievements, potential ratings are forecasts of a worker's future performance and contributions to the firm, making them fundamentally more subjective.

Beyond their use at our firm, Nine Box and similar assessments of potential are nearly ubiquitous in large organizations, where they play a major role in determining promotions, developmental opportunities, and compensation.[3] Before proceeding, it is important to clarify that our goal is not estimate the causal impact of using Nine Box on labor outcomes. Rather, detailed Nine Box data offer a window into how managers think about employee potential more generally. Even if they do not formally use Nine Box, firms and institutions are likely to implicitly consider forecasts of employee potential when making promotion decisions.[4]

Our paper has three main sets of findings. First, women receive lower potential ratings and higher performance ratings than men. The gender gap in potential ratings accounts for up to half of the overall gender gap in promotions. Second, potential ratings systematically understate the future contributions of women: comparing men and women with the same current period potential ratings, women have higher future performance scores and are more likely to remain with the firm. Despite this, firms continue to rate women as having lower potential than men in the following year. In fact, we provide suggestive evidence that these gaps are partly driven by retention motives: men are more likely to exit, especially when they have been passed over for a promotion. Rather than viewing attrition risk as a negative signal of a worker's future contributions to the firm, potential ratings appear to reward men for having less attachment to the firm. Taken together, we provide evidence of misallocation in promotions: marginally promoted women perform better than marginally promoted men. Finally, we consider potential remedies: changing the managers who provide ratings or changing the ratings themselves. For the former, we find no clear evidence that women would benefit from

[3]Practitioners we spoke to at Accenture, Bristol Myers Squibb, CitiGroup, eBay, Intel, JP Morgan, Honeywell, 3M, Ecolab, General Mills, the University of Minnesota, DaVita, and Yale University confirmed that Nine Box or its simpler analogue, Four Box, is widely known and used, including at their own organizations. Cappelli and Keller (2014), Church et al. (2015), and SHRM (2018) discuss the history of Nine Box, its widespread adoption, its applications to succession planning, and related assessments of employee potential.

[4]For example, hiring and promotion committees for professors at most universities do not use the Nine Box system. Yet members of these committees make statements such as, "I think she is better than her dissertation," or, "Her job market paper is carefully done, but I wonder if she has the depth to go after the big questions." Such statements imply separate assessments of a candidate's potential and past performance.

being rated by female or more highly rated managers. For the latter, we identify a trade-off between information and equity: potential ratings, while biased, are nevertheless informative about future performance. Rather than abandoning potential assessments altogether, firms may want to invest in organizational solutions aimed at reducing managers' biases.

We motivate our analysis by documenting a substantial decline in female representation as workers climb the career ladder. In our firm, women constitute 56% of entry level field workers, but only 48% of department managers, 35% of store managers, and 14% of district managers. These patterns are consistent with the "glass ceiling" effect, whereby gendered barriers to promotion compound and yield diminishing shares of women in senior jobs (Blau and Kahn, 2017). Because salaries are closely tied to job levels, gender differences in job levels account for approximately 70% of the overall gender wage gap in our data. This result echoes Petersen and Saporta (2004), which finds that the gender wage gap in the United States largely arises from the assignment of jobs rather than wage discrimination within jobs.

Consistent with declining female representation along the career ladder, we find a robust gender gap in promotions at the individual level: women's annual promotion rates are 10.99% versus 12.62% for men. This baseline gender gap in promotions cannot be explained by differences in past performance: women receive higher performance ratings on average and are 7.4% more likely earn the top performance rating than men.

We show that the gender gap in promotions is better explained by differences in forecasts of potential. Compared to men, women are 12% more likely to earn the lowest potential rating, and 15% and 28% less likely to earn the middle and highest potential ratings, respectively. We also find that potential ratings strongly predict promotions. A one point increase in potential ratings predicts a greater increase in the probability of promotion than a one point increase in performance ratings. Taken together, gender differences in potential ratings can explain up to half of the overall gender promotion gap.

Practitioners describe potential as an individual's ability to contribute to the firm in the future, either through improved performance and greater responsibilities in her original job role or through leadership in a new managerial role (Cappelli and Keller, 2014; Groysberg and Nohria, 2011; Silzer and Church, 2009; Yarnall and Lucy, 2015). Women's lower potential ratings may therefore be justified if they have lower future performance. We show that this is not the case: relative to men with the same current-period Nine Box ratings for performance and potential, women earn higher performance ratings in their next evaluation. In other words, women are more likely to outperform prior ratings of their potential. This result holds whether we consider the worker's future performance in the same role or after promotion to a higher-level role. We further show that managers do not update their evaluations of women's potential upon observing that women outperform men with

similar prior potential ratings. Rather, women continue receiving lower potential ratings in their next evaluation.

One may be concerned that performance ratings do not fully capture a worker's ability to contribute to the firm in the future. In particular, a worker cannot contribute to the firm if she leaves the firm. If women are more likely to leave the firm, then their contributions may be lower even if they outperform men conditional on remaining employed. We next consider whether women may receive lower potential ratings because they have higher rates of attrition or leave-taking due, for instance, to greater care-taking responsibilities (Tô, 2018; England et al., 2016). On the contrary, we find that women are significantly less likely to leave the firm than men. While it is the case that women are more likely to take a leave of absence, the absolute levels of leave among women are too low to explain the large gender gap in potential ratings and promotions.

Interestingly, we find suggestive evidence of a different story: men receive higher potential ratings precisely because they are more likely to leave. Specifically, we show that men are more likely to leave the firm when they are passed over for promotions, or when they receive lower potential ratings, consistent with evidence on gender gaps in outside offers from Blackaby, Booth and Frank (2005).[5] Managers appear to be aware of this, accurately assessing male subordinates to be at greater "risk of loss." Rather than viewing attrition risk as a sign that a worker is less likely to contribute to the firm in the future, managers appear to reward at-risk workers (typically men) with higher potential ratings, translating into higher future pay and greater promotion likelihood. Taken together, men receive higher potential ratings despite being less likely to contribute to the firm by remaining with the firm, and have lower average future performance conditional on staying.

Next, we examine several remaining potential concerns with the interpretation of our results. First, as primary caregivers of children, women may prefer not to take on job roles that entail long hours, low flexibility, or a family relocation to headquarters (e.g. Bertrand, Goldin and Katz, 2010). Thus, women may turn down some promotion offers. While we do not directly observe offers of promotion, we provide suggestive evidence that our results are not well explained by this possibility. Specifically, we continue to find large gender gaps in subsamples where promotions are less likely to conflict with childcare priorities: among older workers who are less likely to have young children at home, and for promotions that do not involve a change in geographic location or a large increase in responsibilities.

Second, performance ratings may not be completely objective, and managers may give female subordinates higher performance ratings than they deserve in order to raise their morale or to compensate for their lower potential ratings. We view this channel as improbable for several reasons.

[5]These results are also consistent with researching showing that men are more likely to complain in response to perceived unfair treatment: for example, Li and Zafar (2022) show that who examine regrade requests in a university setting, suggest that women are also less likely to complain.

Primarily, Nine Box ratings are not disclosed to the employees being rated, so giving women higher performance ratings would not raise morale unless those higher ratings translated into higher compensation or promotion for women, which does not occur in our data. Next, other research (e.g. Sarsons, 2017*b*; Sarsons et al., 2021; Cziraki and Robertson, 2021) suggests that performance ratings, to the extent that they are subjective, are negatively biased against women. We also show that men and women experience similar returns to earning higher performance ratings in terms of promotion; if women's performance ratings were inflated, then we would expect them to have lower returns to performance. Finally, using data from a subsample of sales workers for whom we observe detailed measures of sales, we provide corroborating evidence that women outperform men on an objective metric.

Third, we can only compare future performance among those that stay at the firm. In our data, the composition of men who remain is positively selected: low-performing men have higher attrition than both high-performing men and low-performing women. This pattern, therefore, biases us away from finding that women subsequently outperform men.

Taken together, our main results show that potential ratings matter for promotions and pay, and that firms incorrectly assess women as having too little potential. This raises the question of whether firms promote too few women. Using an IV approach based on variation in the availability of promotion opportunities over time and across roles, we also find that marginally promoted women have higher future performance ratings, relative to marginally promoted men. This finding is indicative of misallocation: our firm could improve the performance of its managers by favoring women on the margin. Our analysis of promotion thresholds also offers an explanation for why women outperform men on average within our sample: if women are held to a higher standard in order to reach or leave each level of the organizational hierarchy, then the set of women in any given level will be positively selected.

Finally, we consider two potential ways of reducing the adverse consequences that potential ratings have on women's careers: changes in the managers that provide ratings, and changes the ratings themselves.

We first ask whether firms can help female employees by changing their manager assignment. It is commonly suggested that women's outcomes can be improved by assigning them to either female or "star" managers, under the logic that female managers may be more effective in mentoring female subordinates and high-performing managers may be better able to evaluate their subordinates in an unbiased manner. We find that gender gaps in potential, pay, and promotions are smaller under female managers. Female managers, however, are also associated with lower overall potential ratings, pay, and promotion rates for their subordinates of both genders, suggesting that female managers are either assigned to weaker subordinates or are tougher in their assessments. Conversely, we find

that gender gaps in potential, pay, and promotions are larger under more highly-rated managers, but that the overall levels of these outcomes are also higher under more highly-rated managers. Taken together, the opposing level and interaction effects in both cases suggest that women are not clearly better off working under either female managers or highly-rated managers.

We next ask whether firms can improve promotion outcomes by varying how potential ratings are used or assigned. We consider two benchmark counterfactuals: ignoring potential ratings and gender altogether, and "de-biasing" potential ratings by making gender-specific adjustments. We show that ignoring potential and gender would nearly eliminate the gender promotion gap, but would also decrease the average future performance of workers who are selected to be promoted. This result reflects the fact that potential ratings do contain useful information about future performance. As a result, promotions based solely on past performance may be less subjective but are also less efficient.

Our second counterfactual increases the potential ratings of women who receive the highest performance rating. This approach, which retains information on both gender and potential, eliminates the gender promotion gap while also increasing the predicted future performance of promoted workers. While this particular policy may be challenging to implement (managers may respond by shading female potential ratings down in anticipation of this gender-specific bonus), it suggests that firms stand to gain from finding ways to de-bias their otherwise informative assessments of potential.

2 Background

2.1 Setting

Our data come from the U.S. operations of a large retailer from February 2009 to October 2015 and contain data on over one million workers, primarily in entry-level hourly roles (e.g. cashiers, sales, customer support, and material handling). Our analysis focuses on the firm's core salaried, full-time employees, spread across the firm's core retail operations and corporate headquarters.

Employees in our firm's corporate headquarters perform a variety of professional functions, the largest of which are information technology, supply chain management, finance, human resources, and real estate management. Career ladders follow a traditional system of pay grades nested within bands. Approximately forty percent of corporate workers with Nine Box ratings are categorized as individual contributors, forty percent are managers, and twenty percent are directors and executives. Although workers receive regular raises, large raises ultimately require workers to be promoted.

Employees in our firm's direct retail operations perform supervisory roles at one of over 4,000 establishments. District managers oversee the functioning of all stores of a given format within their geographically assigned region. They are tasked with reviewing data, setting performance goals, and

making higher-level personnel decisions for the stores they cover. Store managers are tasked with achieving performance goals set out by their district managers but are otherwise given wide latitude in how to achieve them. Their primary activities include analyzing data, formulating store-level strategy, and inspiring their subordinates to successfully execute that strategy. Finally a team of department managers and assistant department managers are responsible for efficiently executing strategies set out by their store manager. This includes customer-facing duties and the hiring and coaching of entry-level staff.

2.2 Nine Box evaluations

Employees in our data are evaluated using a Nine Box grid, a widespread talent assessment and succession planning tool that instructs supervisors to categorize their subordinates into one of nine boxes representing the interaction of high, medium, low ratings on two dimensions: prior job performance and future potential.

The performance dimension of Nine Box ratings is a backward-looking assessment of workers' achievements in their current roles. For instance, store managers may be evaluated on whether their departments met sales targets, replenishment managers may be evaluated on meeting inventory level and delivery targets, and loss prevention managers may be evaluated on inventory lost to theft or damage. In contrast, the potential dimension of Nine Box is a forward-looking assessment. While there is little formal guidance or agreement for what constitutes potential, practitioner guides define potential as a worker's capacity to grow and contribute to the firm in the future within the same role or within the same organization in a different role (Silzer and Church, 2009).

At our firm, Nine Box ratings are assigned annually in a two-step process. First, managers provide initial ratings for their direct management-track subordinates. Managers are not given explicit quotas or curves. Second, there is a district-level calibration meeting during which ratings may be adjusted to ensure that similar standards are being applied. We only observe ratings after the calibration meetings, although our data provider has stated that ratings are rarely adjusted. Final Nine Box ratings serve as the starting point for our organization's annual succession planning process. For example, a vacant position's manager will often reach out to HR to get a list of candidates with strong Nine Box ratings that should be considered for the vacancy. The firm does not post Nine Box ratings publicly or share them with the employees being rated, though interviewed managers note that some supervisors may privately disclose individual ratings. Our discussions with practitioners suggest our firm's procedure for aggregating Nine Box ratings and use of ratings for succession planning and training is very typical of large firms in retail and other industries. Interviewed practitioners also suggest that firms rarely publicize Nine Box ratings due to morale and equity concerns.

In principle, Nine Box allows organizations to distinguish star individual contributors from the best candidates for promotion: a distinction that may be particularly relevant in when the skills necessary to succeed in one role differ from the skills to succeed in another (Baker, Jensen and Murphy, 1988). Otherwise, promoting on prior performance alone can yield substantial mismatch between a worker's skills and their role (Benson, Li and Shue, 2019). Critics, however, argue that Nine Box is less transparent, objective, and consistent than the formal psychometric and skills evaluations that they replaced. In their review of talent management practices, Cappelli and Keller (2014, page 315) summarize:

> "The conceptual idea behind assessing potential has been to identify abilities, given knowledge and skills that presumably can be learned through the development process....however, employers appear to have fallen back on the basic approach of simply asking supervisors to make an assessment of potential, an approach built in to performance appraisals through the Nine Box grid, again made famous by GE. It is a matrix in which performance is assessed on one axis and potential on the other. However, the lack of a definition for what constitutes potential, both within firms and within the academic literature (Groysberg and Nohria, 2011; Silzer and Church, 2009), gives us little reason to believe that this process should produce valid information, despite its widespread use."

Interviews conducted by Yarnall and Lucy (2015) found that even raters themselves believe the Nine Box potential ratings they assigned to be highly subjective. In particular, because supervisors are often provided with limited guidance, ambiguous criteria, and little or no concrete evidence, Nine Box ratings may be prone to well-documented rater biases (Bertrand, Chugh and Mullainathan, 2005). Raters, for instance, may refer to prior years' potential ratings (anchoring bias), first impressions (primacy bias), last impressions (recency bias), and ratings of other dimensions (halo bias) (for a review, see Kahneman, 2011).

Despite these concerns, Nine Box remains a highly popular method of identifying candidates for developmental opportunities and promotion, both because it's easy to implement on its own and integrated into leading HR software packages. As an article in HR Magazine points out: "What's not to like about the Nine Box grid? It's free, easy to use, and ubiquitous." We are not aware of any systematic studies of Nine Box's adoption, and reviews by (Cappelli and Keller, 2014; Cascio and Aguinis, 2008) explicitly conclude that academic work has been conspicuously inattentive to practitioners' widespread use of Nine Box and other so-called talent management practices. However, Nine Box is integrated in major human capital management software packages including Workday, SAP, PeopleSoft, Cezanne, Trakstar, Pipefy, emPerform, which facilitate Nine Box reporting and its translation into development and succession planning. Nine Box Excel templates are also freely available online. Some organizations have also used Nine Box and similar ratings for compensation.

For instance, Microsoft tied performance ratings to cash bonuses and potential ratings to stock options and promotions (Bartlett, 2001).

2.3 Data and summary statistics

We obtain data on Nine Box ratings, promotions, and various demographic characteristics for 29,809 management-track workers employed between 2011 and 2015. These represent the near universe of full-time, salaried, management-track employees at our sample firm during this period. Our data includes workers employed in the firm's corporate headquarters, as well as workers employed across over 4,000 retail locations. Our main data are at the worker-month level.

Nine Box assessments are finalized and recorded in the fourth quarter of each fiscal year, which ends in January. The exact month in which each worker is assessed a new Nine Box rating varies across workers and from year to year. We set a worker's Nine Box rating in each year-month equal to the finalized rating she receives at the end of the relevant fiscal year. Figure 1 shows the labels used by our data provider to describe each box within the Nine Box system. Our data provider reserves the upper left box, representing low performance and high potential, for new hires. Because this rating is mechanically assigned based on tenure, we drop these observations from our analysis sample.

In addition to Nine Box ratings, we observe the following individual-level information: gender, race, tenure in the firm, compensation, job role, subordinates, and manager. For those employed in retail operations, we also observe identifiers for store location.

We determine promotions using data on standardized job titles and annual salary. Most job titles are clearly hierarchical, e.g., a typical career ladder in retail operations can be ordered as assistant department manager, department manager, assistant store manager, store manager, assistant district manager, district manager, vice president, senior vice president, etc. In other cases, the ranking is less clear (e.g., coordinator versus supervisor). We rank job titles by average compensation and classify a worker as having been promoted if, in the next month, we observe a change in job title that is associated with an increase in average compensation associated with that job title or if we observe a change in job title that is associated with a personal raise in salary exceeding 5%. Examples of promotions in our data include moving from department manager to store manager, or moving from web developer to lead web developer.

Table 1 Panel A provides an overview of our sample coverage in terms of workers, time period, and promotion events. Panel B provides summary statistics associated with our sample and key variables. 41% of employees in our sample are female and the average annualized promotion rate is 11.9% (equal to the monthly promotion rate $\times$ 12). Panel C provides pairwise correlations between some of our key variables. Many of paper's empirical results can be previewed in these raw

correlations: being female is positively correlated with performance ratings and negatively correlated with promotion, annual salary, and potential ratings.

In our empirical analysis, we will estimate both the overall gender gap and the conditional gender gap after accounting for a range of control variables. Our goal is not to estimate a "pure" gender effect that is distinct from other correlates of gender such family and cultural background, risk preferences, education, and other demographic factors. Indeed, it is not clear what a pure gender effect would represent, and conditioning on correlated factors that drive gender differences may be a form of over-controlling. To offer a comprehensive view of the data, most of our specifications present both an overall gender gap controlling only for time fixed effects, as well as a gender gap controlling for observable demographics (age, tenure, and race) and establishment location fixed effects. The latter specifications show the remaining gender gap after taking out gender differences that arise due to correlated demographics and location assignment. In the Appendix, we also report our main results comparing promotion and potential ratings for men and women within more narrowly defined job roles: among male and female workers sharing the same manager, same job title, or same pay decile. While our main findings show that gender gaps continue to exist within these narrowly defined roles, we do not include these specifications in our main analysis because one of the the key conclusions of our paper is that differences in job assignment are endogenous to gender; controlling for these variables thereby ignores key selection channels by which gender differences emerge.

3 Main Results

3.1 Gender and promotion

In this section, we begin by describing promotion rates, both in the raw data and controlling for various worker characteristics. In our firm, as in many others, women's representation progressively decreases as one ascends the career ladder, as illustrated in Figure 2. In the left panel, we focus on workers in retail operations, for which there exists a clear ordering of job titles. In stores, women make up 56% of entry level workers[6] (such as cashiers, merchandisers, backroom associates, and salespeople), 48% of department manages, 35% of store managers, and only 14% of district managers. In the right panel, we examine female representation by pay decile (sorted within fiscal year) for all workers with Nine Box ratings within the whole organization. We see a similar pattern of decreasing female representation as one advances in pay deciles. 49% of workers in the bottom pay decile who receive Nine Box ratings are women, compared with 29% at the top.

[6]Note that our main data exclude entry level workers because these workers are generally not salaried or evaluated using the Nine Box. We include them here to describe the overall composition of workers in this firm.

Declining female representation toward the top of the organizational hierarchy is suggestive of a gender gap in promotions to higher level job roles. We explore whether women are less likely to be promoted using the following regression:

$$\text{Promotion}_{it} = a_1 \text{Female}_i + a_2 X_{it} + \delta_y + \varepsilon_{it}. \tag{1}$$

In Equation (1), the level of observation is at the worker-year-month level, where i indexes individuals and t index time measured in months. The sample consists of all full-time workers with Nine Box ratings (these workers are considered management track and exclude entry-level workers such as cashiers). The main outcome of interest is Promotion_{it}, an indicator for whether a worker is promoted in the next month, but we also consider other outcomes such as compensation. Monthly promotion rates are low, so we convert it to an annualized percent by multiplying it by 1,200 (12 months $\times$ 100 percent). The key independent variable is an indicator for whether the worker is female. In all specifications, we control for year fixed effects δ_y to account for time trends. In some specifications, we also controls for a worker's Nine Box performance and/or potential rating, log age, log tenure, race fixed effects, and location fixed effects. Without these control variables, the coefficient on Female_i measures the overall, unconditional gender gap. With these control variables, the coefficient on Female_i measures the unexplained gender gap after accounting for gender differences in control variables X_{it}. Standard errors are clustered by worker to account for account for correlated errors within worker over time.

Table 2 documents a substantial and robust gender gap in promotion rates. Column 1 presents the overall gender gap in our data. The coefficient, -1.64, on the female indicator implies that the annual promotion rate is 1.64% lower for women, or that women are 13.7% less likely to be promoted related to the overall average promotion rate 11.9%. Because this difference in promotion could be due to differences in performance, we control for fixed effects in a worker's Nine Box performance ratings in Column 2 (the omitted category is a performance rating of 1). We find that higher performance ratings are strongly predictive of promotion. More importantly, controlling for worker performance actually increases the gender gap in promotions. As we shall see in future analysis, this occurs because female workers receive higher performance ratings. Once we condition on workers who receive the same performance ratings, we observe a gender promotion gap of -1.84 percentage points, or 15.4%.

In Column 3, we show that part of the gender gap in promotions can be explained by differences in correlated demographic variables. As shown in Table 1 Panel C, women tend to be older and have longer tenure within the firm; these demographic variables are also associated with lower promotion rates. However, even after controlling for these demographic characteristics, women are

1.08 percentage points less likely to be promoted each year (or 9.03% less likely to be promoted relative to the base rate).[7] In Column 4, we find a similarly-sized gender gap after controlling for location fixed effects.

Table 3 documents how differences in promotion rates may lead to differences in compensation. Column 1 shows the overall gender wage gap in our data: the coefficient of -0.118 implies that women's salaries are 12.5% lower than men's. This gap shrinks dramatically to just 3.7% in Column 2, after we control for job level by year fixed effects. Thus, hierarchical differences in assigned job roles account for 70% of the gender wage gap. In Columns 3 and 4, we introduce additional controls for performance and potential ratings, as well as demographic variables and location fixed effects. We continue to find that differences in job levels, which are determined by promotions, appear to be the main determinant of the gender wage gap.

3.2 Gender and potential

We now examine gender gap in evaluations of potential, and show that it can explain a substantial portion of the overall gender promotion gap. Figure 3 plots the gender difference in performance and potential ratings. The left panel plots the distribution of performance and potential ratings for men in our sample, while the right panel represents the differences in shares for women relative to men. In Panel A, we see that men's performance ratings cluster in the middle, with 70% of men receiving a rating of 2, 23% receiving a top rating of 3 and only 7% receiving the lowest rating of 1. In contrast, potential ratings cluster around the lowest ratings, with almost 60% of men receiving a rating of 1, 35% receiving a rating of 2, and only 5% of men receiving a top potential rating.

Compared to the men, Panel B shows that women have higher performance: Women are 7.4% more likely to earn the top performance rating and 21% less likely to receive the lowest performance rating. The opposite pattern, however, occurs for potential ratings. Compared to men, women are 12% more likely to earn the lowest potential rating, and 15% and 28% less likely to earn the medium and high potential ratings, respectively. Thus, women are significantly less likely to earn both the middle and top potential ratings, both of which are valuable because they are relatively rare.[8]

Table 4 documents similar gender differences in Nine Box performance and potential ratings using regression analysis. Panel A shows that women receive substantially higher performance ratings and lower potential ratings. Panel B shows that women are more likely to earn the top performance rating and significantly less likely to earn the top potential rating. These patterns hold both overall and conditional on demographics and location. The divergence in potential and performance ratings

[7] We do not observe significant interaction effects between gender and other demographic characteristics (the "double jeopardy" hypothesis) within our sample.

[8] See Appendix Figure A1 for the raw frequency of observations by gender and the promotion rate within each of the Nine Boxes.

for women is all the more surprising because the two ratings are positively correlated in the overall sample, as shown in Table 1 Panel B.[9] This divergence suggests that potential ratings may be biased against women, a question we evaluate in more detail in Section 3.3.

In Table 5, we examine the extent to which the gender gap in potential ratings can explain the gender gap in promotion. We replicate each column of Table 2, adding controls for a worker's potential rating. By comparing the coefficient on the female indicator in each column of Table 5 with the corresponding coefficient in Table 2, we can estimate the fraction of the gender gap in promotion rates that is explained by gender differences in potential ratings. We find that the coefficient on the female indicator shrinks substantially once we control for potential ratings. Indeed, 53% of the overall gender gap in promotions can be explained by potential ratings. Potential ratings can also explain 48% of the promotion gap conditional on performance ratings, 46% of the promotion gap conditional on performance ratings and demographic characteristics, and 33% of the promotion gap conditional on the above variables and location assignment.

The high explanatory power of potential ratings for the gender promotion gap can be attributed to two forces. First, as seen previously, women are assigned lower potential ratings both unconditionally, and conditional on performance ratings, demographics, and location assignment. Second, Table 5 shows that potential ratings are strong predictors of promotion. In all specifications, we find that a one point increase in potential ratings corresponds to a greater jump in the probability of promotion than a comparable one point increase in the performance ratings. For example, Column 2 shows that a change in potential ratings from 2 to 3 corresponds to a 8.98 percentage point increase in the annual promotion rate, a 75% increase from the baseline in our sample, while a similar change in performance ratings from 2 to 3 corresponds to only a 3.24 percentage point increase in the promotion rate, or a 27% increase from the baseline.

The remaining unexplained gender promotion gap, as measured by the coefficient on the female indicator, in Table 5, may capture several omitted factors. First, women may be less likely to seek or accept advancement opportunities (Fernandez and Mors, 2008; Fernandez-Mateo and Fernandez, 2016). However, we caution that these gender differences in career aspirations may arise endogenously in response to gender bias, and should not necessarily be considered a distinct force. Recent studies have consistently found that stated aspirations are endogenous to perceived opportunities (see, e.g. Correll, 2004). Similarly, Azmat, Cuñat and Henry (2020) find that female lawyers who faced harassment and discrimination report lowered aspirations. Differences in training, particularly related to career development, could also be an omitted factor, though this too may be endogenous. Nine Box potential ratings at our firm (and by convention) are often used to allocate scarce internal

[9] Note that a positive correlation of 0.088 between potential and performance ratings is considered substantial given that these are ordinal variables taking on integer values between 1 and 3.

developmental opportunities; women may be less likely to seek out these opportunities if they anticipate biases in how their potential may be assessed (Milgrom and Oster, 1987).

In addition to our main analysis of the relation between potential ratings and promotions in Table 5, we also provide a supplementary Kitagawa-Oaxaca-Blinder three-fold decomposition in Appendix Table A1. In Panel A, we report the results of an interacted model in which promotion is regression on the female indicator, ratings indicators, and the interaction between female and ratings. In Panel B, we report the decomposition, i.e., the portion of the overall gender gap in promotion rates that can be attributed to differences in the endowments of potential and performance ratings, differences in coefficients (differences in the return to potential and performance ratings), and interactions between endowments and coefficients. The decomposition reveals an overall gender gap in promotion rates of 1.64 percentage points, of which 0.9 (or 55%) can be explained by gender differences in the endowments of potential ratings and -0.16 (or -9.7%) can be explained by differences in the endowments of performance ratings (this figure is negative because women earn higher performance ratings on average). We do not estimate significant or consistently-signed differences in the *returns* to earning higher performance or potential for men and women.

3.3 Information and bias in potential ratings

So far, we have shown that potential ratings help explain why women are less likely to be promoted. In this section, we examine the information content of potential assessments and whether they accurately forecast gender differences in future contributions to the firm. Specifically, we ask: do women receive lower potential ratings because they, in fact, have lower potential?

While the exact definition of "potential" is often debated even within organizations, most practitioners agree that potential ratings should forecast an individual's ability to contribute to the firm in the future, either through improved performance and greater responsibilities in her original job role or through leadership in a new managerial role (Cappelli and Keller, 2014; Groysberg and Nohria, 2011; Silzer and Church, 2009; Yarnall and Lucy, 2015). Thus, effective potential ratings should predict actual future performance, particularly among the sample of workers who are promoted into management positions. We therefore think of workers' future performance ratings as a measure of their "realized potential."

Our analysis focuses on realized potential as defined by performance in the next year. This time horizon is broadly consistent with the practitioner literature and with the views of our data provider, both of which suggest that "future" is typically understood to represent the next few years. We acknowledge, however, that managers within our firm may deviate from this convention (they would face no direct consequences for doing so) and define potential differently, e.g., as a worker's probability of making contributions in the far right tail of the performance distribution, or as a

worker's potential of becoming CEO twenty years in the future. Because of limited sample size and time period, we are not able to assess the informativeness of potential ratings relative to such extremal definitions of future performance, although we note that such definitions are also less likely to be useful for the types of routine job assignments (e.g., promotion from assistant department manager to department manager) that rely on Nine Box ratings.

Table 6 Panel A shows that high current potential ratings predict higher performance ratings 12 months in the future. Our estimates in Column 1 indicate that, relative to those with a low potential rating (the median for the sample), workers with a high potential rating have a 0.17 point higher performance rating in the following fiscal year (equivalent to a 1/4 of a standard deviation change in performance). Importantly, this positive and significant correlation holds even after conditioning on the worker's current Nine Box performance rating. That is, potential ratings appear to contain real information about a worker's future performance, beyond what can be forecast using information on prior performance alone.[10] This correlation holds in both the full sample of workers, as well as within the subsample of employees who experience a promotion event (so that their future performance ratings reflect performance in a new role).

Since potential ratings predict future performance, one natural explanation for why women may receive lower potential ratings is that they are likely to have worse future performance.[11] If women's lower potential ratings are indeed justified, then we would expect that, controlling for current period potential ratings, men and women should have similar future performance ratings. Assessments of potential may be biased against women if, for the same potential rating, women have systematically higher measures of realized potential.

Table 6 Panel A relates a worker's current period potential ratings with their next period performance rating (measured 12 months in the future). Columns 1 and 2 focus on the full sample of workers, where "next period" performance can refer to either performance in the same role or in a different role. Column 1 controls for year fixed effects while Column 2 also controls for location fixed effects and demographics. In both cases, we find that, controlling for a worker's current potential and performance ratings, women receive higher future performance ratings than their male colleagues. That is, women systematically outperform forecasts of their potential. In Columns 3 and 4, we limit the sample to workers promoted in the current year-month and again regress future performance ratings on a female indicator and pre-promotion ratings. Since the sample of promoted workers is much smaller, and some locations only have one promotion event within our sample period, we

[10] This is consistent with Li (2017), which shows that ignoring advice from biased advisers would reduce the overall quality of investment decisions, because biased advice still contains useful signals of a project's quality.

[11] Azmat and Ferrer (2017) find that differences in billable hours and new business origination explain about half of the gender gap in lawyers' pay. Relatedly, Cook et al. (2018) find female Uber drivers earn less per hour than men despite identical pay contracts due to differences in experience and driving preferences. However, Sarsons (2017*a*) finds female surgeons have better performance than male surgeons.

exclude location fixed effects in this and all other analysis restricted to the promoted subsample. We again find a similarly-sized significant positive coefficient on the female indicator, implying that promoted women outperform promoted men, conditional on current potential and performance ratings and other observable control variables.

3.4 Updating beliefs about potential

We next consider how firms update evaluations of women's potential in response to observing their future performance. To do this, we replicate the previous analysis, using 12-month ahead *potential* ratings as the outcome of interest. Panel B of Table 6 shows that women continue to receive significantly lower future potential ratings compared to men with the same current performance and potential ratings, both in the full sample and in the sample of newly promoted workers. This pattern is all the more noteworthy because performance and potential ratings are determined simultaneously by the same managers. This means that, at the same time that women are given performance ratings indicating that they outperformed their previous year's potential ratings (relative to men with the same potential ratings), women are still assessed as having lower potential going forward.

4 Interpretation and robustness

In this section, we present a variety of results that further explore our main findings.

We begin by showing that variation in the gender potential gap relates to geographic variation in gender-related attitudes. In particular, our firm maintains establishments across the United States, which we are able to link to county-level measures of gender inequality. We focus on the labor market components of the World Bank Gender Inequality Index: female representation in management-level positions, gender wage gaps, and female educational attainment (full details for the construction of the county-level measures are provided in the Appendix). Role congruity theory predicts that managers in areas with lower gender equity may find it more difficult to imagine women succeeding in management positions because they do not frequently observe women in leadership roles. Consistent with this, Appendix Table A2 shows that managers rate women as having lower potential (relative to men with the same performance ratings) in counties with lower female representation in management-level positions, larger gender wage gaps, and lower female educational attainment These patterns support the view that perceptions and stereotypes may play a role in limiting women's assessed potential. We acknowledge, however, that we lack exogenous variation in county-level gender inequality measures, so these results should be viewed as suggestive.

Next, we address several possible alternative explanations for our results. While we have thus far considered demand side drivers of women's lower promotion rates, another possibility is that women

are simply less likely to accept promotions, if offered. Indeed, past research has shown that women are more likely to be primary caregivers of children. Faced with childcare responsibilities, women may prefer not to accept job roles that entail long hours, lower flexibility, or a family relocation to headquarters (e.g. Bertrand, Goldin and Katz, 2010; Goldin, 2014; Goldin and Katz, 2016).

Although we do not directly observe offers of promotion, we can reproduce our main tests in subsamples where accepting a promotion offer would be less likely to conflict with childcare priorities. In Appendix Table A3, we continue to find large gender gaps among workers over the age of 50, who are less likely to have young children at home. In Appendix Table A4, we explore gender gaps in promotion depending on whether the promotion requires an out-of-state relocation. Here, we uncover interesting and nuanced results. We find a large gender gap in out-of-state promotions, of which only a small fraction can be explained by gender differences in potential ratings: relocation costs, rather than potential ratings, appear to play a dominant role in constraining women's access to out-of-state promotions. Yet, when we examine in-state promotions, which constitute 90% of promotions in our data, we find a different story. In this sample, the overall gender promotion gap is smaller (10%, relative to 15% in our full sample), but a much greater proportion of it can be explained by gender differences in potential ratings—nearly 80%.[12] Taken together, these results suggest that women face multiple barriers to advancement. When a promotion requires relocation, women may be held back by household considerations; when a promotion does not require relocation, women may instead be held back by perceptions of their potential.

A second alternative interpretation is that, rather than potential ratings being downward biased for women, performance ratings are upward biased for women. Although performance can be directly observed and are thus inherently less subjective than potential ratings, performance ratings are nonetheless unlikely to be completely objective. One possible concern for our analysis that is managers may give women higher performance ratings than they deserve in order to raise their morale or to compensate women for their lower potential ratings. We believe this channel is improbable because Nine Box ratings are not disclosed to the individuals being rated: women would not know if they had higher performance ratings unless these ratings translated into higher compensation or promotion (women in our data have lower pay and promotion rates). In addition, related research (e.g. Sarsons, 2017*b*; Sarsons et al., 2021; Cziraki and Robertson, 2021) shows that, to the extent that performance assessments are subjective, they tend to be negatively biased against women.

[12] We estimate gender gaps for out-of-state and same-state promotions by dividing the coefficient for the female indicator by the dependent variable mean within each sample to estimate the extent to which women are less likely to be promoted relative to the sample mean. Women are 62% less likely to experience an out-of-state promotion and 10% less likely to experience a same-state promotion. We estimate the fraction of the gender gap explained by potential ratings using the change in the coefficient on the female indicator after controlling for potential ratings. We find that 12% of the out-of-state promotion gap can be explained by potential ratings and 77% of the same-state promotion gap can be explained by potential ratings.

We also provide empirical evidence that women's performance ratings are unlikely to be inflated in our data. First, we consider the subsample of our data for which we observe an objective measure of job performance: credited sales to sales workers. In Appendix Table A5, we find that female sales worker sell approximately six percent more than male sales workers. Female sales workers outperform both overall, and after conditioning on demographics and detailed location-month fixed effects to account for seasonality in retail sales. While the sales worker sample does not overlap with our main sample of employees with Nine Box ratings (sales workers are not considered management-track), these results show that women outperform men in another large group of workers at our firm. Returning to our main sample of management-track workers, we also test a prediction of the hypothesis that women get higher performance ratings than they deserve: if managers inflate women's performance ratings to increase their morale, their inflated performance ratings should translate relatively less into higher promotion rates for women. In our Kitagawa-Oaxaca-Blinder decomposition, presented in Appendix Table A1, we show that this is not the case: men and women experience similar returns to performance ratings, in terms of their promotion probabilities.

A third possibility is that our results are driven by selection into the sample of workers who remain with the firm (for whom we observe measures of future performance). If high-performing men are relatively more likely to exit, then our results showing that women have higher future performance may only be valid among the set of workers who stay.[13] We explore this possibility by examining turnover by gender, ratings, and their interaction in Appendix Figure A2 and Appendix Table A6. Among both men and women, workers who remain at the firm tend to be more positively selected on performance, and this relationship is, if anything, stronger for men. As a result, attrition does not remove the strongest men from our sample: rather, remaining women outperform remaining men despite the latter being slightly more positively selected.

Our final set of analysis in this section show that our main results remain similar under several alternative specifications. So far, our analysis has shown that women receive: (a) higher performance ratings, (b) lower potential ratings, (c) lower promotion rates, and (d) higher future performance ratings, and (e) lower future potential ratings. We consider a series of robustness checks for these five main results.

For our main analysis, we purposely do not control for job roles because differences in job assignment are endogenous to gender. Nevertheless, in Tables A7, A8, and A9, we find that gender gaps persist even after controlling for proxies of job roles such as manager fixed effects, job level fixed effects, and pay decile fixed effects, respectively.

[13] We note that firms define potential as a worker's future contributions. Under this definition, a worker who leaves should be counted as having zero contributions to the firm. In the next section, we show that men have higher rates of attrition. As such, women contribute more to their employers by having higher performance when they stay, and by being less likely to contribute zero by leaving.

Our next set of results consider robustness to various technical specifications. In our main analysis, we use a monthly panel because promotions, salary changes, and Nine Box ratings are updated in a staggered fashion throughout the year. In A10, we find similar gender gaps using data collapsed to an annual panel. In A11, we cluster standard errors by manager instead of by worker and continue to estimate highly statistically significant gender gaps of the same magnitude. Finally, for ease of exposition, our main analysis controlled for potential ratings and performance ratings separately. Appendix Table A12 presents a fully interacted model with indicators for every possible Nine Box combination of potential and performance ratings, and finds similar results.

5 Leaves of absence, retention, and risk of loss

Our results in Section 3.3 show that women's lower potential ratings cannot be justified by weaker future performance, conditional on remaining with the firm. Potential ratings, however, may also reflect managers' expectations about a worker's commitment to the firm. For example, a large literature has shown that women are more likely to leave the workforce after having children, and those who remain often experience stagnation in wage growth and greater disruptions to their productivity (e.g. Bertrand, Goldin and Katz, 2010; Kleven, Landais and Sogaard, 2019; Cubas, Juhn and Silos, 2021).[14] If managers believe that women's careers are more likely to be interrupted or cut short by family care duties, they may lower their assessments of women's potential. In this section, we explore this question empirically, leaving aside the legality or ethics of such behavior.

In Table 7 we examine attrition from the firm entirely. Column 1 demonstrates that, on average, women have lower attrition than men: women's lower potential ratings can not be justified by concerns that they are more likely to exit the firm. In Column 2, we consider whether workers are more likely to exit when they are "passed over" for a promotion, which we code as having occurred if another worker reporting to the same manager is promoted (moves to a higher position in the next month) while the focal worker is not promoted. We find that men who are passed over are 32% more likely to exit the firm, relative to the baseline rate of exit; among women who are passed over (with the same Nine Box ratings), this figure is only 12%. This difference in willingness to exit the firm is even more pronounced among high performers. In Columns 3-4, we repeat this exercise for workers who receive the highest Nine Box performance rating. Among this group, men who are passed over are 40% more likely to leave relative to the base exit rate, whereas women in the same position are at most only 1.5% more likely to leave.

The fact that men are at higher risk of attrition may impact how they are treated by the firm. In Table 8, we consider the relation between gender, perceptions of attrition risk, and potential

[14] Indeed, recent work shows that women may themselves underestimate the labor market impacts of having children (Kuziemko et al., 2018).

ratings. For three years of our data, we observe firm ratings for each employee's "risk of loss," a three point rating capturing a worker's risk of leaving the firm. In Column 1, we show that risk of loss ratings are indeed predictive of future attrition: workers rated as being at high risk of loss are 61% more likely to exit the firm, relative to those at low risk. In Column 2, we see that women receive substantially lower risk of loss ratings, relative to men with the same Nine Box performance and potential ratings. Finally, Columns 3 through 6 show how perceptions of attrition risk may help explain why men achieve better outcomes along a range of dimensions. In Column 3, we see that risk of loss ratings are positively and significantly related to a worker's next potential rating (measured 12 months in the future), controlling for current performance and potential ratings. In Columns 4 and 5, we find that higher risk of loss ratings are also associated with significantly higher promotion probability and compensation. Finally, we note that the coefficient on female remains large and negative throughout these regressions, implying that the gender gap cannot be completely explained by women's lower risk of loss and potential ratings.

Taken together, our results suggest that firms anticipate men's higher rates of attrition in their risk of loss assessments, and react by granting men higher next period potential ratings, promotions, and pay. This reaction could operate through two complementary channels. First, managers may infer (possibly incorrectly) that workers who talk about outside offers and hint about leaving truly have higher potential. Second, managers may directly seek to retain high risk-of-loss subordinates by giving them higher potential ratings (which would translate into higher promotion probability). This latter channel may be related to the agency problem of "talent hoarding" (Friebel and Raith, 2013; Haegele, 2021), in which self-interested managers seek to keep their best subordinates instead of promoting them. If men are relatively more likely to leave when passed over for promotion, managers may prefer to hoard their female subordinates.

Regardless of the exact channel, the positive relation between risk of loss and potential ratings and promotions implies that firms essentially reward the threat of exit, rather than perceiving it as a negative signal of a worker's commitment or ability to contribute to the firm in the future. Yet, as can be seen in Tables 6 and 7, this leads the firm to be more likely to promote men who, relative to their female peers with the same Nine Box ratings, tend to have lower future performance and higher rates of future attrition.

These results also suggest that, if firms are concerned about retention, they could improve outcomes by employing more women. Because women have greater attachment to the firm, this would reduce the extent to which the firm needs to sacrifice managerial match quality in promotion decisions to retain workers. Further, because women also have higher performance and lower wages, such a policy is unlikely to lead to lower worker productivity or an increased wage bill.

Finally, in Table 9, we consider differences in the probability men or women take a leave of absence, defined as temporary time off of work that could be paid or unpaid. The most common reasons for taking a leave of absence are related to family and child care, or personal or family medical issues. Columns 1 shows that women, indeed, are substantially more likely to take a leave of absence from the firm. The coefficient on the female indicator implies that women are 0.45 percentage points more likely to be on leave in the following month, 65% higher than the baseline of 0.70 percentage points. In Columns 2-4, we explore how this difference relates to women's potential ratings. Columns 2 reports the raw gender potential gap restricted to the slightly reduced sample for which we observe leaves of absence data. Columns 3 shows that the gender gap in potential ratings remains similar in magnitude after controlling for the worker's past leaves measured in number of months. Column 3 shows that the gender gap in potential ratings remains similar even after controlling for realizations of leaves in the future. In both cases, past and future leaves are negatively related to potential ratings, but the gender gap in potential ratings appears to exist separately from inferences about leave.

Of course, managers may assign female subordinates lower potential ratings because of expectations about future leaves, and we would not be able to control for these expectations using only data on actual future leaves, as in Column 4. We can instead conduct the following thought experiment: how much extra future leave must managers believe women will take to explain the gender gap in potential ratings? Column 3 implies that, based on the relationship between past leaves and potential ratings, managers would have to believe that women take on average four extra months of leave to justify a gender gap of 0.086 points in potential. These beliefs do not match the data: compared to men, women take an extra 0.05 months of leave per year relative to men. Even if the manager considers potential leaves over the next 10 years, women on average only take an additional half of month of leave relative to men. In other words, while women take relatively more leave than men, their absolute levels of leave are too low to explain the large gender gap in potential ratings that we observe.

6 Misallocation in promotions

So far, our results have demonstrated that potential ratings inaccurately reflect the future contributions of women in our firm: comparing men and women with the same current period Nine Box ratings, women contribute more both through stronger performance in the future and through lower attrition. We have also shown that potential ratings play a large role in the firm's promotion decisions. This raises the following question: if firms are under-rating the potential of their female employees, does this mean that they are also under-promoting them?

In this section, we conduct a Becker outcomes test for discrimination in firm *promotion* decisions. If the firm holds women to a higher promotion standard, then it could also increase the quality of its managers by promoting more women on the margin.

Following Benson, Li and Shue (2019), we identify "marginally promoted" applicants using an instrument for promotions (described shortly). In our application, which also builds on prior work by Abadie (2003) and Arnold, Dobbie and Yang (2018), the instrument is not used to identify a causal effect, but rather to identify a set of instrument compilers. Intuitively, compliers to a promotion instrument can be thought of as marginal: they are promoted if they receive a good draw of the instrument, but not otherwise. We therefore compare the realized potential (e.g. future performance ratings) of male and female instrument compliers. If marginally promoted women outperform marginally promoted men, then the firm requires women to meet a higher standard for promotion.

We instrument worker i's promotion outcome at time t using Z_{it}, the average promotion rate for workers with the same job title in the same year t, leaving out all workers in worker i's same office or store location. Similar to Benson, Li and Shue (2019), this promotion instrument captures the idea that workers employed during employment expansions are more likely to be promoted irrespective of their performance or potential.

A natural concern with this instrument is that employment expansion may be correlated with future Nine Box ratings: for example, instrument compliers promoted in expansions may face more favorable circumstances and may be credited with higher performance as a result. We address this concern in several ways. First, in our analysis, we measure a worker's future performance rating residualized for job title interacted with year fixed effects. That is, we consider a worker's future performance relative to other workers with the same job in the same year: by construction, this measure of realized future performance is not related to job-time level changes, such as changing consumer demand, that may play a role in shaping our promotion rate instrument. Another potential concern is reverse causality: if a given worker is particularly strong, the firm may chose to promote her, generating a higher promotion rate for that worker's job title at that time. Using a jackknife approach and leaving out a worker's own promotion status (and that of her colleagues in the same office or store) severs the correlation between our instrument and an individual worker's quality. Finally, note that we are ultimately interested in the *difference* in future performance ratings for marginally promoted men and women. As such, we are only concerned about biases in our instrument that differ for men and women; our analysis is unbiased as long as male and female IV compliers are promoted into similar economic conditions.

To compute the future ratings of compliers, we estimate the following regressions:

$$Y_{it} \times P_{it} = \alpha_0 + \alpha_1 P_{it} + X'_{it}\alpha + \varepsilon_{it} \text{ if female} \quad (2)$$

$$Y_{it} \times P_{it} = \beta_0 + \beta_1 P_{it} + X'_{it}\beta + \nu_{it} \text{ if male} \quad (3)$$

In Equations (2) and (3), $Y_{it} \times P_{it}$ is worker i's future rating outcome Y_{it} (either future performance score or future potential score) if she is promoted ($P_{it} = 1$) and to zero otherwise. We include controls for performance ratings, year and, in some specifications, demographics and location. The OLS coefficients $\hat{\alpha}_1^{OLS}$ and $\hat{\beta}_1^{OLS}$ estimates average future performance ratings among all promoted women and men, respectively, after controlling for other covariates. The IV estimates $\hat{\alpha}_1^{IV}$ and $\hat{\beta}_1^{IV}$, in contrast, represent future ratings among female and male compliers, respectively. This logic is analogous to the idea that IV estimates identify a local average treatment effect among compliers.[15] Our analysis focuses on the difference between female and male workers who are promoted on the margin: $\hat{\alpha}_1^{IV} - \hat{\beta}_1^{IV}$.

Figure 4 presents the results of our analysis, with the accompanying regressions reported in Appendix Table A13. In Panel A, we see that marginally promoted women have higher future performance ratings, relative to marginally promoted men.

This finding suggests that gender bias in potential ratings leads to misallocation in managerial opportunities. To see this explicitly, consider the following modification to the firm's existing promotion policy P:

$$\widetilde{P}(X) = \begin{cases} P(X)^{Z=1} \text{ if female,} \\ P(X)^{Z=0} \text{ if male.} \end{cases}$$

This new promotion policy modifies the firm's existing practices by favoring women on the margin. Specifically, consider two workers, male and female, with the same covariates X. The new policy $\widetilde{P}$ asks that women be evaluated for promotion as if the firm had many promotion opportunities available ($Z = 1$), and men be evaluated as if the firm had few such opportunities ($Z = 0$).[16] By construction, the promotion decisions of this new policy differ only in its treatment of instrument compliers. In particular, women who would have been promoted only when overall promotion rates were high would always be promoted under this new policy, but men in the same situation would not be. The difference in future managerial performance between $\widetilde{P}$ and P is therefore given by the difference in the future performance of female and male compliers. As demonstrated in Panel A of

[15] For a detailed proof, see Benson, Li and Shue (2019).

[16] For simplicity in exposition, we let Z be a binary instrument in this example (whether job level promotion rates are above or below median) though in practice we use a continuous variable.

Figure 4, this difference is positive: the firm can improve the quality of its managers by favoring women on the margin.

Panel B of Figure 4 repeats our IV analysis with future potential ratings as the outcome of interest, rather than future performance ratings. We find that the same women who have higher future performance than their male peers continue to receive lower potential ratings going forward. These results echo our earlier findings from Table 6 showing that the firm does not update its evaluations of women's potential upon observing their future performance.

Finally, Appendix Table A13 reports that these IV results hold when controlling for demographics and location. To provide additional context, this table also compares the future performance ratings of the average and marginal promoted worker. For both men and women, the average promoted worker performs better in the future than the marginal promoted worker. Taken together, our findings suggest that firms are more likely to promote higher quality workers relative to lower quality workers within gender, but that women appear to be held to a higher threshold.

7 Potential policy responses

In this section, we consider two possible HR policy responses: changes to the managers making Nine Box ratings decisions, and changes to Nine Box ratings practices themselves. We note that we do not have random assignment of managers to subordinates, nor do we observe random variation in Nine Box scoring policy, so the results in this section are meant to be suggestive.

7.1 Heterogeneity by manager assignment

We begin by documenting how gender gaps in ratings, pay, and promotions vary across different types of managers. In particular, we focus on two manager characteristics: gender and the manager's own Nine Box performance and potential ratings. This analysis is motivated by the common suggestion that women would benefit from working under female managers, who may be less biased against other women and act as mentors and advocates for their female subordinates. McGinn and Milkman (2013), for instance, show that female managers serve as role models for their female subordinates, and enhance their career progression.[17] Likewise, women may benefit from working under higher quality managers who may be better at assessing their subordinates' true performance and potential or less likely to hoard their talented subordinates.

Throughout this analysis, we regress a worker's performance and potential rating, pay, or promotion outcomes on gender, the manager characteristic of interest (gender or Nine Box rating),

[17] However, it is not obvious that female subordinates would be better off working under a female manager. Research on the "queen bee" syndrome shows that female managers can be tougher on their female subordinates viewed as competition (see e.g., Lee et al. (2015)).

and the interaction between worker gender and manager characteristics. We caution, however, that we cannot distinguish between treatment or selection effects. That is, subordinate outcomes may differ across managers both because managers are assigned to different types of subordinates and because managers differ in how they assess or advocate for their subordinates.

In Table 10, we examine whether subordinates' ratings depend on their own gender and the gender of the manager who is rating them, and is motivated by studies that have found such interaction effects on termination and career advancement (e.g., Egan, Matvos and Seru, 2017; Cullen and Perez-Truglia, 2020). We find that gender gaps in potential ratings and pay are smaller (but not eliminated) under female managers. However, we also find that female managers are associated with lower overall levels of ratings, pay, and promotion rates for their subordinates, regardless of subordinate gender.[18] A female employee can therefore expect a smaller gender *gap*, but not necessarily an increase in the absolute levels of ratings, pay, or promotion rates. These opposing level and interaction effects echo related results in Cardoso and Winter-Ebmer (2010), which shows that female-led firms are associated with lower gender wage gaps as well as lower levels of wages.

In Table 10, we explore how worker outcomes vary with their manager's Nine Box performance and potential evaluations. Our results mirror those for manager gender: subordinates assigned to managers with higher performance and potential ratings have higher levels of ratings, pay, and promotion, but gender gaps in these outcomes are also larger. On net, it is unclear whether women benefit from these manager assignments.[19]

7.2 Counterfactual promotion policies

In this section, we consider the impact of counterfactual promotion policies on both equity, as measured by differences in promotion rates for men and women, and on efficiency, as measured by the expected future performance of promoted candidates.

We consider two counterfactual policies. The first counterfactual we consider is simply to stop using gender and potential in promotion decisions. The second policy we consider is to continue using potential ratings, but to first adjust them to account for gender bias. We accomplish this in a simple way: we increase the potential ratings of women with the highest performance ratings (e.g. among women with top performance scores, those with "low" potential ratings are now rated

[18] Our conversations with practitioners reveal a possible explanation for why female managers give lower ratings on average to their subordinates. While most firms do not set quotas for Nine Box ratings, firms do provide guidance that managers should hold workers to a high or tough standard. If female managers are more conscientious about following such guidelines, such behavior would translate into lower ratings under female managers.

[19] A possible explanation for these patterns is that managers who themselves receive higher performance and potential ratings are stronger advocates for their subordinates. These managers give higher ratings for subordinates on average, and the higher average allows for greater variation in ratings across subordinates, magnifying the gender gap.

"medium," those rated "medium" are now rated "high," and those who were already rated high keep the same potential rating).

To assess the impact of these counterfactual policies, we begin by estimating a regression of promotion on gender, demographics, and fixed effects for potential ratings, performance ratings, and year. The fitted values from this regression represent the firm's baseline promotion policy. To evaluate the impact of the first counterfactual—blinding promotion to potential and gender information—we form estimates of promotion likelihood by setting the coefficients on gender and potential to zero. To evaluate the impact of the second counterfactual policy—increasing potential scores for high performing women—we simply use our adjusted potential ratings as inputs into predicting promotion and form new fitted values.

To form estimates of the counterfactual gender gap in promotions under each policy, we report average fitted promotion likelihoods for men and for women. To form estimates of the quality of each counterfactual promotion policy, we report averages of workers' next period performance ratings, weighted by workers' fitted likelihood of promotion under each counterfactual policy. That is, if a worker is more likely to be promoted under a given policy, we place more weight on this worker's expected future performance. Our analysis computes average future performance ratings in two ways: over the full sample of workers and the sample of workers who are actually promoted. Averaging future performance in the full sample offers more complete coverage, but risks conflating a worker's observed performance in their actual role rather than in the role our counterfactual policy would assign them to. To address this concern, we also limit our analysis to the subsample of workers who are actually promoted in order to ensure that our next-period performance ratings reflect true performance in the promoted roles.[20]

We report the results of this exercise in Table 11. Columns 1 and 2 compare promotion rates for men and women under each policy, while Columns 3-5 present estimates of the quality of promoted workers under each policy. Focusing on Columns 1 and 2 in the first two rows, we show that blinding promotion decisions to gender and potential ratings substantially reduces the gender gap in promotions by 65%, from a 1.7 percentage point gap to a 0.6 percentage point gap. In Columns 3 and 4 we show that this increase in equity comes at an efficiency cost: estimated next period performance of promoted workers decline relative to the baseline promotion policy. This is true when applying promotion likelihood weights to full sample of workers, as well as the subsample of promoted workers. These results are consistent with Table 6, which shows that, despite being biased, potential ratings do contain useful information about future performance. A promotion policy that ignores potential ratings would therefore discard important information about future productivity.

[20]We note that this approach can introduce some selection bias: if women are positively selected into promotion relative to men, then our analysis may overstate the true performance of counterfactually promoted women by excluding the performance of women who are not actually promoted in practice.

In the third row of Table 11, we consider what happens when we retain information on potential ratings, but apply adjustments aimed at "undoing" gender bias. We find that a targeted shift in potential ratings, applying only to women who are rated highest in terms of performance, leads to an improvement in both equity and efficiency. This approach eliminates the gender promotion gap, while also increasing the estimated next period performance of promoted workers.

Before continuing, we note that our conclusions are subject to two caveats. First, our analysis in Column 4 restricts to workers who are actually promoted. If the set of women who are promoted by our "de-biased" counterfactual policy—but not actually promoted in practice—are differentially weaker than the set of men in this position, then our findings could overstate the efficacy gains of this policy be excluding non-promoted workers.[21] Second, the de-biasing policy we consider is easy to circumvent: managers can simply shade women's ratings downward in anticipation of them receiving a gender-specific bonus. Given this, we regard our counterfactual not as a specific policy proposal, but as a demonstration that firms may be able to increase both the quality and equity of their promotion decisions by identifying ways to retain the information content of potential ratings, while addressing the level effects of bias.

8 Conclusion

We show that the widely-used practice of forecasting workers' "potential" as a basis for allocating training and job assignments contributes to gender gaps in promotion and pay. Despite receiving higher performance ratings, women are assessed as having lower potential. These lower potential ratings can explain up half of the observed gender gap in promotions.

Women's lower potential ratings may be justified if women indeed contribute less to the firm in the future. Our findings, however, indicate that this is not the case. Among employees with the same current performance and potential ratings, women outperform men on evaluations of their future performance and are less likely to exit the firm. These mistakes in potential ratings do not appear to self-correct: even though women outperform their potential ratings, they continue to receive lower potential evaluations in the future. These persistently low potential ratings apply regardless of whether woman continue in their current roles, or are promoted and perform well in their new roles.

We find that addressing bias in potential ratings using commonly-discussed approaches may be challenging. First, one cannot simply decrease the gender promotion gap by having more female managers. The presence of female managers attenuates the potential and promotions gap to some extent but, on net, female managers still give lower potential ratings to women. This suggests that policies that seek to decrease the gap between assessed potential and future performance need to

[21] Of course, if the reverse is true and men are differentially positively selected into promotion, then we would understate the gains of promoting more women.

address broader organizational questions rather than simply changing the gender of the evaluator. Similarly, we find that assigning women to higher quality managers would not reduce gender bias. While managers who themselves receive higher performance and potential ratings appear to be stronger advocates for their subordinates on net–they give them higher ratings and salaries, these benefits accrue almost entirely to male subordinates of high-performing managers: gender gaps in performance, potential, and promotions expand under such managers.

Second, our findings suggest that doing away with potential ratings altogether would reduce gender inequities but at the cost of reducing the quality of promoted managers. A growing empirical literature now supports the long-held anecdotal belief that the best workers do not always make the best managers. When current performance is an imperfect indicator for future performance, it is reasonable for firms to look for other ways of assessing potential. In our data, potential ratings predict future performance even after accounting for current performance; ignoring this information would therefore reduce the quality of the firm's decisions.

Our results instead show that there may be large gains from finding ways to de-bias assessments of potential. One approach is to boost potential ratings and promotion rates for women who are rated as low-potential and high-performing. Such women are are rarely promoted despite their tendency to succeed when they are. An alternative approach would substitute indicators of potential with one that is less prone to stereotypes of who may be an effective leader. In recent years, firms have made various attempts to increase promotions and retention among women and minorities, from the use of bias-conscious algorithms in screening to training programs focused on conscious and unconscious bias. This paper suggests that these would be fruitful areas for further research.

Finally, we emphasize that our results should not be viewed as estimating the causal impact of Nine Box adoption or other similar ratings systems on women's career advancement. Rather, we view Nine Box ratings as providing a quantifiable window into how managers evaluate worker's potential separately from their observed performance. Such assessments, regardless of whether they are formalized into Nine Box ratings, are a critical part of hiring and promotion decisions in many organizations, and may contribute to gender gaps in career advancement more broadly.

References

Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of econometrics*, 113(2): 231–263.

Arnold, David, Will Dobbie, and Crystal S Yang. 2018. "Racial bias in bail decisions." *The Quarterly Journal of Economics*, 133(4): 1885–1932.

Azmat, Ghazala, and Rosa Ferrer. 2017. "Gender gaps in performance: Evidence from young lawyers." *Journal of Political Economy*, 125(5): 1306–1355.

Azmat, Ghazala, Vicente Cuñat, and Emeric Henry. 2020. "Gender promotion gaps: Career aspirations and workplace discrimination." *CEPR Discussion Paper No. DP14311*.

Babcock, Linda, and Sara Laschever. 2009. *Women don't ask*. Princeton University Press.

Babcock, Linda, Maria P Recalde, Lise Vesterlund, and Laurie Weingart. 2017. "Gender differences in accepting and receiving requests for tasks with low promotability." *American Economic Review*, 107(3): 714–47.

Baker, George P., Michael C. Jensen, and Kevin J. Murphy. 1988. "Compensation and Incentives: Practice vs. Theory." *The Journal of Finance*, 43(3): 593–616.

Bartlett, Christopher. 2001. "Microsoft: Competing on Talent (A)." Harvard Business School case study 9-300-001.

Benson, Alan, Danielle Li, and Kelly Shue. 2019. "Promotions and the Peter Principle*." *The Quarterly Journal of Economics*, 134(4): 2085–2134.

Bertrand, Marianne, Claudia Goldin, and Lawrence F Katz. 2010. "Dynamics of the gender gap for young professionals in the financial and corporate sectors." *American economic journal: applied economics*, 2(3): 228–55.

Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan. 2005. "Implicit discrimination." *American Economic Review*, 95(2): 94–98.

Biasi, Barbara, and Heather Sarsons. 2020. "Flexible wages, bargaining, and the gender gap." National Bureau of Economic Research.

Blackaby, David, Alison L Booth, and Jeff Frank. 2005. "Outside offers and the gender pay gap: Empirical evidence from the UK academic labour market." *The Economic Journal*, 115(501): F81–F107.

Blau, Francine D, and Lawrence M Kahn. 2017. "The gender wage gap: Extent, trends, and explanations." *Journal of economic literature*, 55(3): 789–865.

Brands, Raina A., and Isabel Fernandez-Mateo. 2017. "Leaning Out: How Negative Recruitment Experiences Shape Womens Decisions to Compete for Executive Roles." *Administrative Science Quarterly*, 62(3): 405–442.

Bureau, U.S. Census. 2019. https://www.census.gov, Accessed: 2021-05-26.

Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais. 2017. "'Acting Wife': Marriage Market Incentives and Labor Market Investments." *American Economic Review*, 107(11): 3288–3319.

Cappelli, Peter, and JR Keller. 2014. "Talent management: Conceptual approaches and practical challenges." *Annu. Rev. Organ. Psychol. Organ. Behav.*, 1(1): 305–331.

Cardoso, Ana Rute, and Rudolf Winter-Ebmer. 2010. "Female-led firms and gender wage policies." *Industrial and Labor Relations Review*, 64(1): 143–163.

Cascio, Wayne F, and Herman Aguinis. 2008. "Research in industrial and organizational psychology from 1963 to 2007: Changes, choices, and trends." *Journal of Applied Psychology*, 93(5): 1062.

Church, Allan H, Christopher T Rotolo, Nicole M Ginther, and Rebecca Levine. 2015. "How are top companies designing and managing their high-potential programs? A follow-up talent management benchmark study." *Consulting Psychology Journal: Practice and Research*, 67(1): 17.

Cook, Cody, Rebecca Diamond, Jonathan Hall, John A List, and Paul Oyer. 2018. "The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers." National Bureau of Economic Research.

Correll, Shelley J. 2004. "Constraints into preferences: Gender, status, and emerging career aspirations." *American sociological review*, 69(1): 93–113.

Correll, Shelley J, Katherine R Weisshaar, Alison T Wynn, and JoAnne Delfino Wehner. 2020. "Inside the black box of organizational life: The gendered language of performance assessment." *American Sociological Review*, 85(6): 1022–1050.

Cubas, German, Chinhui Juhn, and Pedro Silos. 2021. "Work-Care Balance over the Day and the Gender Wage Gap." *AEA Papers and Proceedings*, 111: 149–53.

Cullen, Zoë B, and Ricardo Perez-Truglia. 2020. "The Old Boys' Club: Schmoozing and the Gender Gap." National Bureau of Economic Research.

Cziraki, Peter, and Adriana Robertson. 2021. "Credentials Matter, but Only for Men: Evidence from the S&P 500." *Available at SSRN 3894730*.

Eagly, Alice H, and Steven J Karau. 2002. "Role congruity theory of prejudice toward female leaders." *Psychological review*, 109(3): 573.

Egan, Mark L, Gregor Matvos, and Amit Seru. 2017. "When Harry fired Sally: The double standard in punishing misconduct." National Bureau of Economic Research.

England, Paula, Jonathan Bearak, Michelle J. Budig, and Melissa J. Hodges. 2016. "Do Highly Paid, Highly Skilled Women Experience the Largest Motherhood Penalty?" *American Sociological Review*, 81(6): 1161–1189.

Fang, Lily Hua, and Sterling Huang. 2017. "Gender and connections among Wall Street analysts." *The Review of Financial Studies*, 30(9): 3305–3335.

Fernandez-Mateo, Isabel, and Roberto M Fernandez. 2016. "Bending the pipeline? Executive search and gender inequality in hiring for top management jobs." *Management Science*, 62(12): 3636–3655.

Fernandez, Roberto M, and Marie Louise Mors. 2008. "Competing for jobs: Labor queues and gender sorting in the hiring process." *Social Science Research*, 37(4): 1061–1080.

Friebel, Guido, and Michael Raith. 2013. "Managers, training, and internal labor markets." *Simon School Working Paper No. FR 13-31*.

Goldin, Claudia. 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review*, 104(4): 1091–1119.

Goldin, Claudia, and Lawrence F Katz. 2016. "A most egalitarian profession: pharmacy and the evolution of a family-friendly occupation." *Journal of Labor Economics*, 34(3): 705–746.

Groysberg, Boris, and Nitin Nohria. 2011. "How to hang on to your high potentials." *Harvard Business Review*, 77–83.

Haegele, Ingrid. 2021. "Talent Hoarding in Organizations." *Working paper*.

Human Development Reports. 2021. http://hdr.undp.org/en/content/gender-inequality-index-gii, Accessed: 2021-05-26.

Kahneman, Daniel. 2011. *Thinking, fast and slow.* Macmillan.

Kaplan, Steven N, Mark M Klebanov, and Morten Sorensen. 2012. "Which CEO characteristics and abilities matter?" *The Journal of Finance*, 67(3): 973–1007.

Kleven, Henrik, Camille Landais, and Jakob Egholt Sogaard. 2019. "Children and Gender Inequality: Evidence from Denmark." *American Economic Journal: Applied Economics*, 11(4): 181–209.

Koenig, Anne M, Alice H Eagly, Abigail A Mitchell, and Tiina Ristikari. 2011. "Are leader stereotypes masculine? A meta-analysis of three research paradigms." *Psychological bulletin*, 137(4): 616.

Kuziemko, Ilyana, Jessica Pan, Jenny Shen, and Ebonya Washington. 2018. "The Mommy Effect: Do Women Anticipate the Employment Effects of Motherhood?" National Bureau of Economic Research Working Paper 24740.

Lee, Sun Young, Marko Pitesa, Stefan Thau, and Madan M Pillutla. 2015. "Discrimination in selection decisions: Integrating stereotype fit and interdependence theories." *Academy of Management Journal*, 58(3): 789–812.

Li, Cher Hsuehhsiang, and Basit Zafar. 2022. "Ask and You Shall Receive? Gender Differences in Regrades in College." *American Economic Journal: Economic Policy.*

Li, Danielle. 2017. "Expertise versus Bias in Evaluation: Evidence from the NIH." *American Economic Journal: Applied Economics*, 9(2): 60–92.

McGinn, Kathleen L., and Katherine L. Milkman. 2013. "Looking Up and Looking Out: Career Mobility Effects of Demographic Similarity Among Professionals." *Organization Science*, 24(4): 1041–60.

Milgrom, Paul, and Sharon Oster. 1987. "Job discrimination, market forces, and the invisibility hypothesis." *The Quarterly Journal of Economics*, 102(3): 453–476.

Milgrom, Paul R. 1988. "Employment contracts, influence activities, and efficient organization design." *Journal of political economy*, 96(1): 42–60.

Peter, Laurence J., and Raymond Hull. 1969. *The Peter Principle.* New York: William Morrow & Co.

Petersen, Trond, and Ishak Saporta. 2004. "The opportunity structure for discrimination." *American Journal of Sociology*, 109(4): 852–901.

Player, Abigail, Georgina Randsley de Moura, Ana C Leite, Dominic Abrams, and Fatima Tresh. 2019. "Overlooked leadership potential: The preference for leadership potential in job candidates who are men vs. women." *Frontiers in psychology*, 10: 755.

Prendergast, Canice, and Robert Topel. 1993. "Discretion and bias in performance evaluation." *European Economic Review*, 37(2-3): 355–365.

Proudfoot, Devon, Aaron C Kay, and Christy Z Koval. 2015. "A gender bias in the attribution of creativity: Archival and experimental evidence for the perceived association between masculinity and creative thinking." *Psychological science*, 26(11): 1751–1761.

Roussille, Nina. 2020. "The central role of the ask gap in gender pay inequality." *URL: https://ninaroussille. github. io/files/Roussille_ askgap. pdf.*

Sarsons, Heather. 2017*a*. "Interpreting signals in the labor market: evidence from medical referrals." *Working paper.*

Sarsons, Heather. 2017*b*. "Recognition for group work: Gender differences in academia." *American Economic Review*, 107(5): 141–45.

Sarsons, Heather, and Guo Xu. 2021. "Confidence Men? Evidence on Confidence and Gender among Top Economists." Vol. 111, 65–68.

Sarsons, Heather, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. 2021. "Gender differences in recognition for group work." *Journal of Political Economy*, 129(1): 101–147.

SHRM. 2018. "Succession Planning: What is a 9-box grid?" https://www.shrm.org/resourcesandtools/tools-and-samples/hr-qa/pages/whatsa9boxgridandhowcananhrdepartmentuseit.aspx, [Online; accessed 29-Nov-2021].

Silzer, Rob, and Allan H Church. 2009. "The pearls and perils of identifying potential." *Industrial and Organizational Psychology*, 2(4): 377–412.

Tô, Linh T. 2018. "The signaling role of parental leave." *Harvard University.*

Yarnall, Jane, and Dan Lucy. 2015. "Is the Nine Box Grid All About Being in the Top Right?" *Roffrey Park research report.*

FIGURE 1: NINE BOX RATINGS AND LABELS

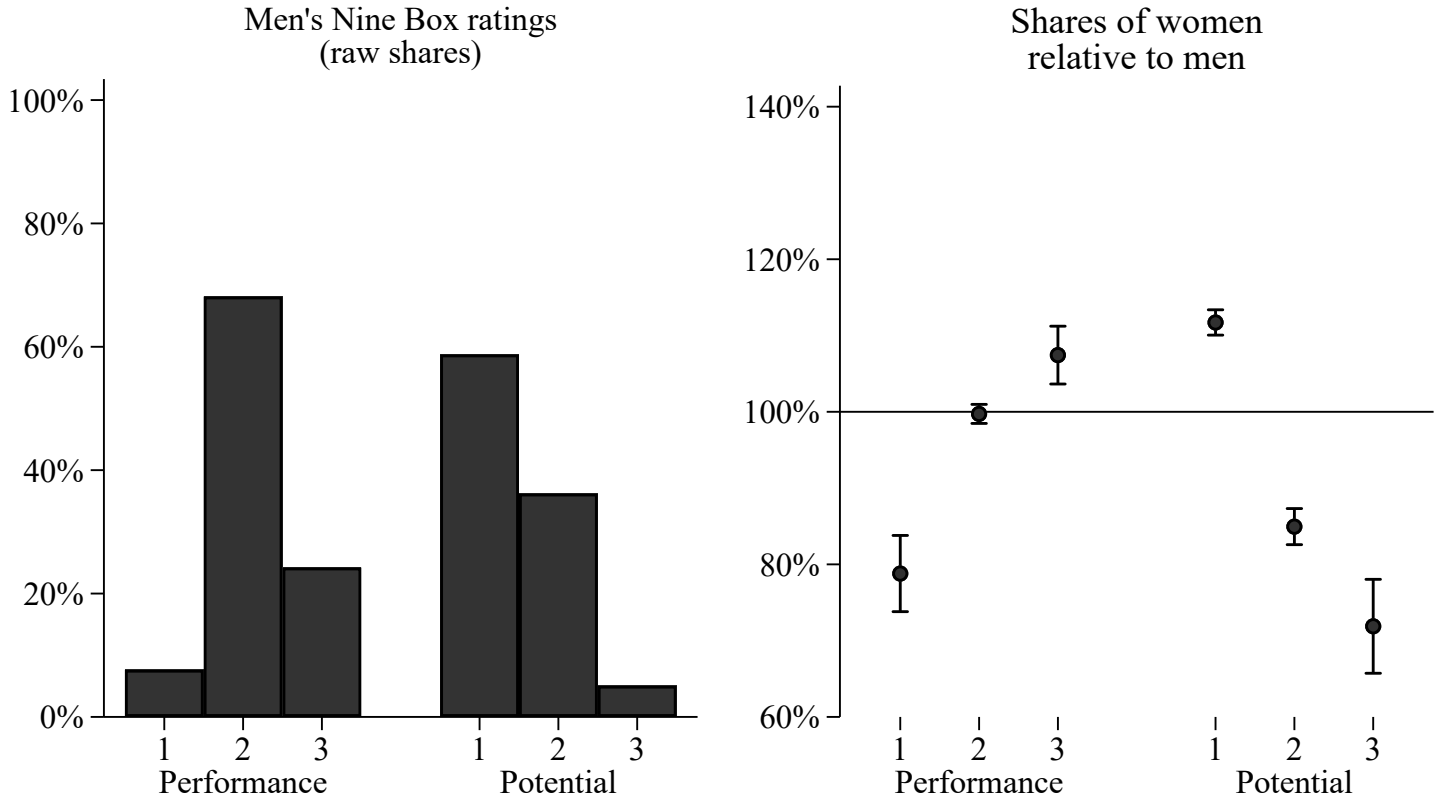
		Performance		
		1 (Low)	2 (Medium)	3 (High)
Potential	3 (High)	New hire	Delivering, strong potential	high-performing, top talent
	2 (Medium)	Potential mismatch	Delivering, promotable	high-performing, promotable
	1 (Low)	Underperforming	Delivering	high-performing, critical resource

NOTES: The table reports facsimiles of the labels used by our data provider. The box for low performance, high potential is explicitly used for new hires.

FIGURE 2: FEMALE SHARES IN THE ORGANIZATIONAL HIERARCHY

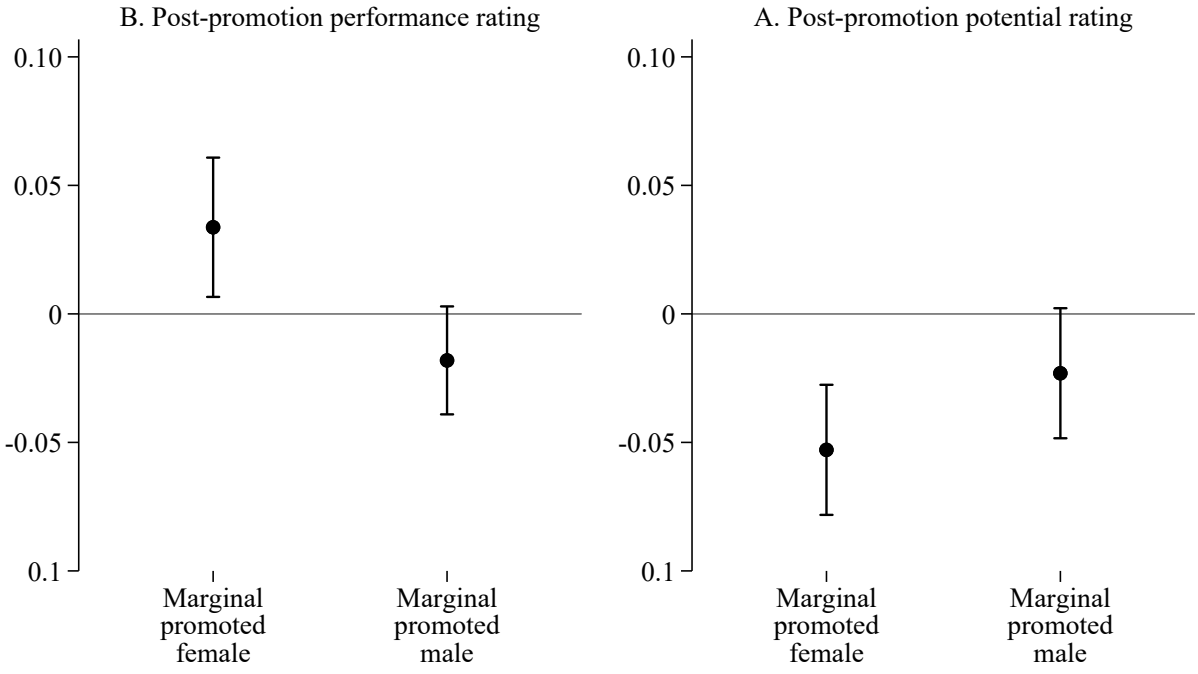
NOTES: The left panel reports the female share among retail operations workers. Department managers include all managers junior to the location's head manager, including associate managers overseeing departments and salaried assistant general managers. Counts include the number of unique workers who held a job at each level within our sample period. The right panel reports the female share among all workers who receive Nine Box ratings. This population includes all regular, salaried workers, including corporate workers and field workers at the level of department managers and above, and excludes entry-level retail workers. The deciles are sorted by regular annual salaries.

FIGURE 3: GENDER GAPS IN NINE BOX RATINGS



NOTES: The left panel represents the distribution of Nine Box performance ratings and potential ratings assigned to male workers. The right panel represents the share of women relative to men who receive the rating in the horizontal axis. Vertical brackets represent 95% confidence intervals.

FIGURE 4: POST-PROMOTION NINE BOX RATINGS FOR MARGINAL PROMOTIONS



NOTES: This figure reports estimates for the IV coefficient on the promoted indicator from Equation (3), as described in the text. Panel A focuses on a promoted worker's 12-month-ahead performance rating, whereas Panel B focuses on that worker's 12-month-ahead potential rating. Vertical brackets represent standard errors.

TABLE 1: SUMMARY STATISTICS AND CORRELATIONS

Table 1: Descriptive statistics

Panel A: Data coverage						
Locations	$> 4,000$		Worker-months			900209
Workers	29809		Promotions			8964
Months (2011-2015)	58					
Panel B: Summary statistics		**mean**	**sd**	**p25**	**p50**	**p75**
(a) Female		.412	.492	0	0	1
(b) Promotion (annualized percent)		11.9	119.148	0	0	0
(c) Salary (annual dollars)		70691	101974	45000	59188	85000
(d) Potential rating		2.18	.536	2	2	2
(e) Performance rating		1.429	.578	1	1	2
(f) Age		44.4	10.834	35.8	45	53
(g) Tenure (months)		171.2	139.534	53	138	253
(h) White		.736	.441	0	1	1
(i) Black		.09	.287	0	0	0
(j) Hispanic		.103	.304	0	0	0
(k) Asian		.058	.234	0	0	0
(l) Other race		.012	.111	0	0	0
Panel C: Correlations	**(a)**	**(b)**	**(c)**	**(d)**	**(e)**	**(f)**
(a) Female	1					
(b) Promotion (annualized percent)	-.007	1				
(c) Salary (annual dollars)*	-.132	-.019	1			
(d) Potential rating	.032	.025	.206	1		
(e) Performance rating	-.071	.048	.176	.088	1	
(f) Age*	.037	-.052	.193	.015	-.271	1
(g) Tenure (months)*	.072	-.039	-.021	.097	-.215	.465

NOTES: Asterisks denote that salary, age, and tenure are computed as log variables in Panel C and subsequent analyses.

Table 2: Gender gap in promotions

Promoted	(1)	(2)	(3)	(4)
Female	-1.644***	-1.837***	-1.027***	-1.079***
	(0.267)	(0.266)	(0.255)	(0.280)
Performance rating				
2=Med		6.498***	6.061***	5.417***
		(0.325)	(0.331)	(0.378)
3=High		11.35***	11.74***	10.99***
		(0.424)	(0.426)	(0.482)
Fiscal year FEs	Yes	Yes	Yes	Yes
Demographic controls			Yes	Yes
Location FEs				Yes
Observations	900209	900209	900209	900209

Notes: This table reports a linear probability model for promotions. The dependent variable takes a value of 1200 if the worker is promoted in the following month, and zero otherwise, so that coefficients represent annualized percents. The omitted category for the performance rating is 1=Low. Demographic controls include log age, log tenure, and race/ethnicity fixed effects for White (omitted category), Black, Asian, Hispanic, and Other. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Table 3: Gender pay gap and the role of promotions

Log salary	(1)	(2)	(3)	(4)
Female	-0.118***	-0.0364***	-0.110***	-0.0416***
	(0.00566)	(0.00391)	(0.00508)	(0.00367)
Potential rating				
2=Med			0.164***	0.0828***
			(0.00402)	(0.00287)
3=High			0.286***	0.136***
			(0.00957)	(0.00633)
Performance rating				
2=Med			0.134***	0.0653***
			(0.00577)	(0.00422)
3=High			0.291***	0.153***
			(0.00719)	(0.00508)
Fiscal year FEs	Yes	Yes	Yes	Yes
Job level × year FEs		Yes		Yes
Demographic controls			Yes	Yes
Observations	899726	899726	899726	899726

Notes: This table reports regressions of log salary on the female indicator, performance rating indicators (the omitted category is 1=Low), potential rating indicators (the omitted category is 1=Low), fiscal year fixed effects, and/or job level interacted with fiscal year fixed effects. Data exclude 87 FLSA-exempt field sales managers working only on commission. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

TABLE 4: GENDER DIFFERENCES IN NINE BOX RATINGS

Panel A	Potential rating		Performance rating	
	(1)	(2)	(3)	(4)
Female	-.083***	-.0527***	.0343***	.0151***
	(.0057)	(.0052)	(.0053)	(.005)
Mean of DV	2.1799	2.1799	1.4288	1.4288
	(.0006)	(.0006)	(.0006)	(.0006)
Fiscal year FEs	Yes	Yes	Yes	Yes
Demographic controls		Yes		Yes
Location FEs		Yes		Yes
Observations	900209	900209	900209	900209
Panel B	**Top potential rating**		**Top performance rating**	
	(5)	(6)	(7)	(8)
Female	-.0143***	-.0137***	.0181***	.0061
	(.0017)	(.0019)	(.0044)	(.0043)
Mean of DV	.2496	.2496	.0446	.0446
	(.0005)	(.0005)	(.0002)	(.0002)
Fiscal year FEs	Yes	Yes	Yes	Yes
Demographic controls		Yes		Yes
Location FEs		Yes		Yes
Observations	900209	900209	900209	900209

NOTES: This table reports regressions of Nine Box performance and potential ratings on the female indicator and other control variables for fiscal year fixed effects, worker demographics, and location fixed effects as described in Table 2. Panel A uses the raw rating (1, 2, or 3) as the dependent variable whereas Panel B uses an indicator for whether the worker received the top performance or potential rating. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

TABLE 5: POTENTIAL AND PROMOTIONS

Promoted	(1)	(2)	(3)	(4)
Female	-0.771***	-0.963***	-0.555**	-0.726***
	(0.256)	(0.256)	(0.256)	(0.279)
Potential rating				
2=Med	10.61***	10.52***	7.343***	6.838***
	(0.294)	(0.292)	(0.282)	(0.299)
3=High	20.70***	19.50***	14.66***	13.57***
	(0.865)	(0.864)	(0.631)	(0.650)
Performance rating				
2=Med		6.856***	6.358***	5.921***
		(0.329)	(0.503)	(0.536)
3=High		10.09***	10.71***	10.38***
		(0.417)	(0.546)	(0.588)
Fiscal year FEs	Yes	Yes	Yes	Yes
Demographic controls			Yes	Yes
Location FEs				Yes
Observations	900209	900209	900209	900209

NOTES: This table replicates Table 2, with the addition of control variables for potential rating indicators (the omitted category is 1=Low). By comparing the coefficient on the female indicator in this table with the corresponding coefficient in Table 2, we estimate the fraction of the gender gap in promotions that can be explained gender differences in potential ratings. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Table 6: Bias in potential ratings and promotions

Panel A	Full sample		Promoted sample	
Next performance rating	(5)	(6)	(7)	(8)
Female	.0328***	.0197***	.0285*	.0279*
	(.0046)	(.0048)	(.0154)	(.0154)
Potential rating				
2=Med	.0913***	.1021***	.0678***	.0665*
	(.0048)	(.0052)	(.0162)	(.0163)
3=High	.1677***	.1974***	.1266***	.1275***
	(.0116)	(.0118)	(.0278)	(.0282)
Performance rating				
2=Med	.3637***	.2613***	.2697***	.2609***
	(.0111)	(.0116)	(.0522)	(.0513)
3=High	.7671***	.5801***	.5139***	.4975***
	(.0121)	(.0126)	(.0534)	(.0524)
Fiscal year FEs	Yes	Yes	Yes	Yes
Demographic controls		Yes		Yes
Location FEs		Yes		
Observations	586338	586338	5222	5222
Panel B	Full sample		Promoted sample	
Next potential rating	(1)	(2)	(3)	(4)
Female	-.0482***	-.0346***	-.0685***	-.0536***
	(.0047)	(.005)	(.018)	(.0175)
Potential rating				
2=Med	.4241***	.2871***	.2461***	.2074***
	(.0055)	(.0059)	(.0186)	(.0184)
3=High	.7297***	.5388***	.4593***	.3816***
	(.0167)	(.0164)	(.0359)	(.0353)
Performance rating				
2=Med	.2135***	.1668***	.1082*	.0775
	(.0091)	(.0096)	(.0592)	(.062)
3=High	.3147***	.2935***	.1795***	.175***
	(.0099)	(.0106)	(.0602)	(.063)
Fiscal year FEs	Yes	Yes	Yes	Yes
Demographic controls		Yes		Yes
Location FEs		Yes		
Observations	586338	586338	5222	5222

Notes: Panel A reports a regression of Nine Box potential ratings 12 months in the future on control variables as described in Table 2. Columns 3 and 4 restrict the sample to worker-months corresponding to promotion events and do not control for location fixed effects due to the smaller sample size. Panel B is identical to Panel A except that the dependent variable is the Nine Box performance rating 12 months in the future. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

TABLE 7: ATTRITION AND GENDER

Attrition	Full sample		High performers	
	(1)	(2)	(3)	(4)
Female	-0.593*	-0.455	0.0586	0.294
	(0.350)	(0.357)	(0.607)	(0.622)
Passed over		8.260***		8.121**
		(1.917)		(3.216)
Female × Passed over		-5.127***		-7.817**
		(1.913)		(3.127)
Fiscal year FEs	Yes	Yes	Yes	Yes
Potential rating FEs	Yes	Yes	Yes	Yes
Performance rating FEs	Yes	Yes		
Observations	886899	886899	221876	221876
DV mean	25.967	25.967	20.292	20.292

NOTES: This table reports regressions of whether a worker leaves the firm in the next month on gender and other measures. Attrition takes values of 0 or 1,200 so that coefficients can be interpreted as annual percents. The variable "Passed over" is equal to one if another worker sharing the same manager is promoted in the next month, but the focal worker is not. Regressions that include this variable also include a control for whether there is a promotion in that month for this team. Columns 1-2 report results for the full sample of workers (excluding the last year-month observation for any given location to avoid right truncation in our panel data). Columns 3-4 repeat this exercise, but for the subset of workers who receive a current period Nine Box performance rating of 3, the top rating. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Table 8: Risk of loss and potential ratings

	(1) Attrition	(2) Risk of loss rating	(3) Next potential	(4) Promoted	(5) Next log salary	(6) Next performance
Risk of loss rating						
2=Med	4.964***		0.0834***	0.587	0.110***	0.00946
	(0.469)		(0.00644)	(0.359)	(0.00581)	(0.00613)
3=High	13.85***		0.0918***	3.892***	0.148***	-0.00702
	(1.038)		(0.0136)	(0.758)	(0.0120)	(0.0127)
Female		-0.0547***	-0.0432***	-0.929***	-0.125***	0.0319***
		(0.00685)	(0.00561)	(0.314)	(0.00605)	(0.00559)
Fiscal year FEs	Yes	Yes	Yes	Yes	Yes	Yes
Potential rating FEs	Yes	Yes	Yes	Yes	Yes	Yes
Performance rating FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	533780	533780	415683	533780	376100	415683
DV mean	22.657	1.421	1.382	10.852	11.093	2.189

Notes: Column 1 reports regressions of actual attrition in the next month on "risk of loss" ratings assigned by the firm and other control variables, where attrition takes values of 0 or 1,200. Risk of loss is categorized by the firm as 1-low, 2-medium, or 3-high. Column 2 regresses the risk of loss rating on the female indicator and other control variables. Columns 3-5 examine the relationship between risk of loss and gender with the 12-month-ahead potential rating, whether a worker is promoted in the following month, and log salary, respectively. See the appendix for a parallel analysis controlling for demographics and location fixed effects. The sample includes all worker months prior to the last month that given location is in our sample, to allow for observations of future behavior. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

TABLE 9: LEAVE OF ABSENCE AND THE GENDER POTENTIAL GAP

	(1) Leave of absence	(2) Potential rating	(3) Potential rating	(4) Potential rating
Female	0.453***	-0.0856***	-0.0840***	-0.0817***
	(0.0358)	(0.00562)	(0.00562)	(0.00562)
Past leaves			-0.0231***	-0.0185***
			(0.00405)	(0.00396)
Future leaves				-0.0269***
				(0.00298)
Fiscal year FEs	Yes	Yes	Yes	Yes
Performance FEs	Yes	Yes	Yes	Yes
Observations	886899	886899	886899	886899

NOTES: Column 1 present a regression of whether a worker takes a leave of absence in the next month on the female indicator and other control variables as described in Table 2. Columns 2-4 relate gender and leaves to Nine Box potential ratings. Past leaves measures the number of months of leave a worker has taken in their past history with the firm and future leaves measures the number of months of leave taken in the future within our data sample. Results with additional demographic and location controls are in the appendix. The sample includes all worker months prior to the last month that given location is in our sample, to allow for observations of future behavior. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

TABLE 10: VARIATION BY MANAGER CHARACTERISTICS

Potential rating	(1)	(2)	(3)	(4)
Female	-0.0582***	-0.0623***	-0.0188	-0.0332***
	(0.00582)	(0.0228)	(0.0152)	(0.0102)
Manager characteristic	-0.0286***	0.00157***	0.0326***	0.0348***
	(0.00732)	(0.000339)	(0.00461)	(0.00397)
Female × Manager characteristic	0.0193**	0.000206	-0.0152**	-0.0119**
	(0.00970)	(0.000474)	(0.00665)	(0.00596)
This model's manager characteristic	Manager female	Manager age	Manager's performance rating (1-3)	Manager's potential rating (1-3)
Fiscal year FEs	Yes	Yes	Yes	Yes
Demographic controls	Yes	Yes	Yes	Yes
Location FEs	Yes	Yes	Yes	Yes
Observations	885353	885353	829429	829429

NOTES: Column 1 examines heterogeneity in the focal worker's potential ratings by the manager's gender. Column 2 examines variation by a linear term for age in years. Including a quadratic term, the estimated Female potential gap peaks under mid-career managers (age 45) and is statistically significant with $p < 0.01$ from the 5th to 95th percentiles of manager ages. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

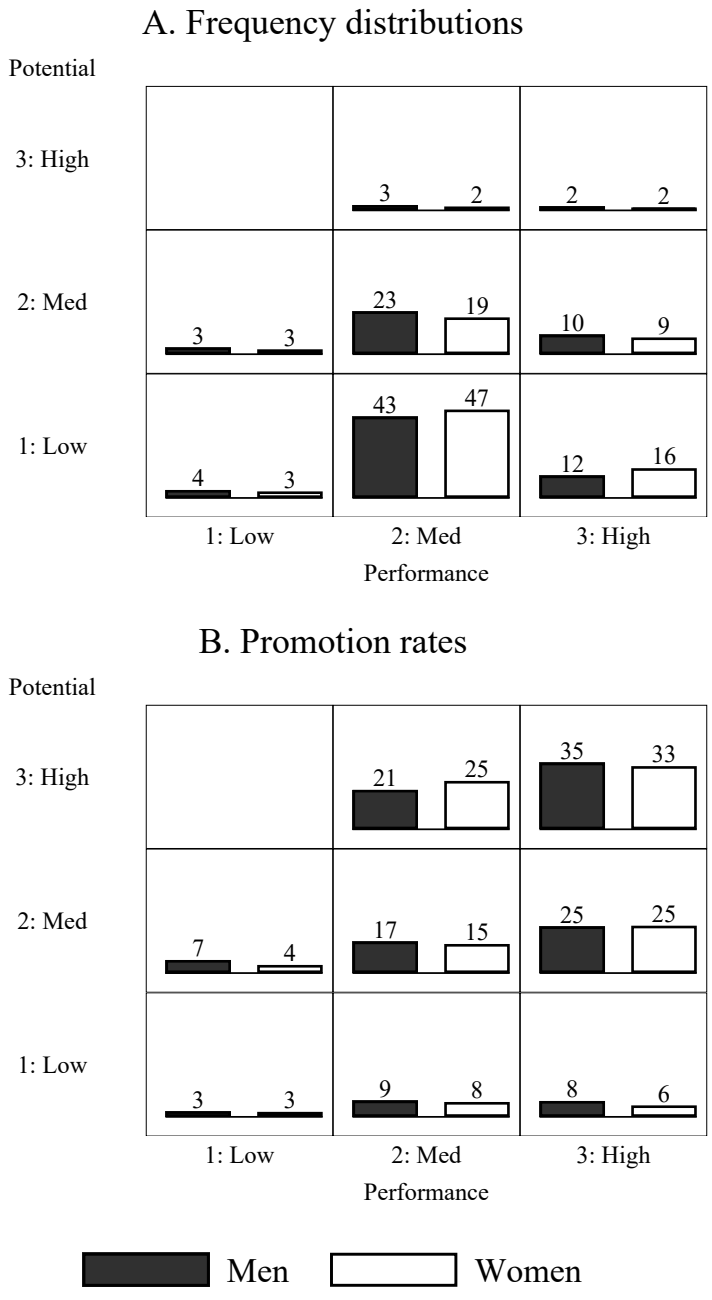
TABLE 11: PROMOTION AND PERFORMANCE UNDER COUNTERFACTUAL POLICIES

	Promotion rates		Expected next performance rating among promoted		
	(1)	(2)	(3)	(4)	(5)
	among men	among women	full sample	true unpromoted sample	true promoted sample
Baseline: current promotion policy	12.6232 (.1815)	10.9882 (.1426)	2.2933 (.0045)	2.2936 (.0045)	2.2743 (.0091)
Counterfactual 1: ignore potential scores and gender	12.2036 (.1492)	11.5865 (.1426)	2.2772 (.0041)	2.2774 (.0041)	2.2633 (.0092)
Counterfactual 2: add one to the potential scores of all women	12.6232 (.1776)	18.0554 (.3682)	2.2797 (.0041)	2.2798 (.0041)	2.2711 (.0091)
Counterfactual 3: add one to the potential scores of high performing women	12.6232 (.1732)	12.7829 (.2233)	2.3111 (.0046)	2.3115 (.0046)	2.2822 (.0093)

NOTES: This table reports expected promotion rates and expected future performance ratings under the firm's current promotion policy and counterfactual promotion policies. Details are provided in Section 7.2. Columns 1 and 2 provide the counterfactual expectations of promotion rates for men and women, respectively. Column 3 provides expectations of the 12-month-ahead performance ratings among the promoted, weighted by the current estimated promotion probabilities for all workers. Column 4 does the same, but for workers who were not promoted in the true sample. Column 5 does the same, but for workers who were promoted in the true sample. The baseline policy uses predicted values of promotion rates based on gender, performance ratings, potential ratings, year, age, tenure, and race/ethnicity. Counterfactual 1 uses the baseline policy, but omits gender and potential ratings when estimating promotion rates. Counterfactual 2 adds one to the potential ratings of women who receive potential ratings of 1 or 2 when estimating predicted promotion rates. Counterfactual 3 adds one to the potential ratings of women who receive potential ratings of 1 or 2, but only for women who receive performance ratings of 3. Bootstrapped standard errors, clustered by worker, are in parentheses.

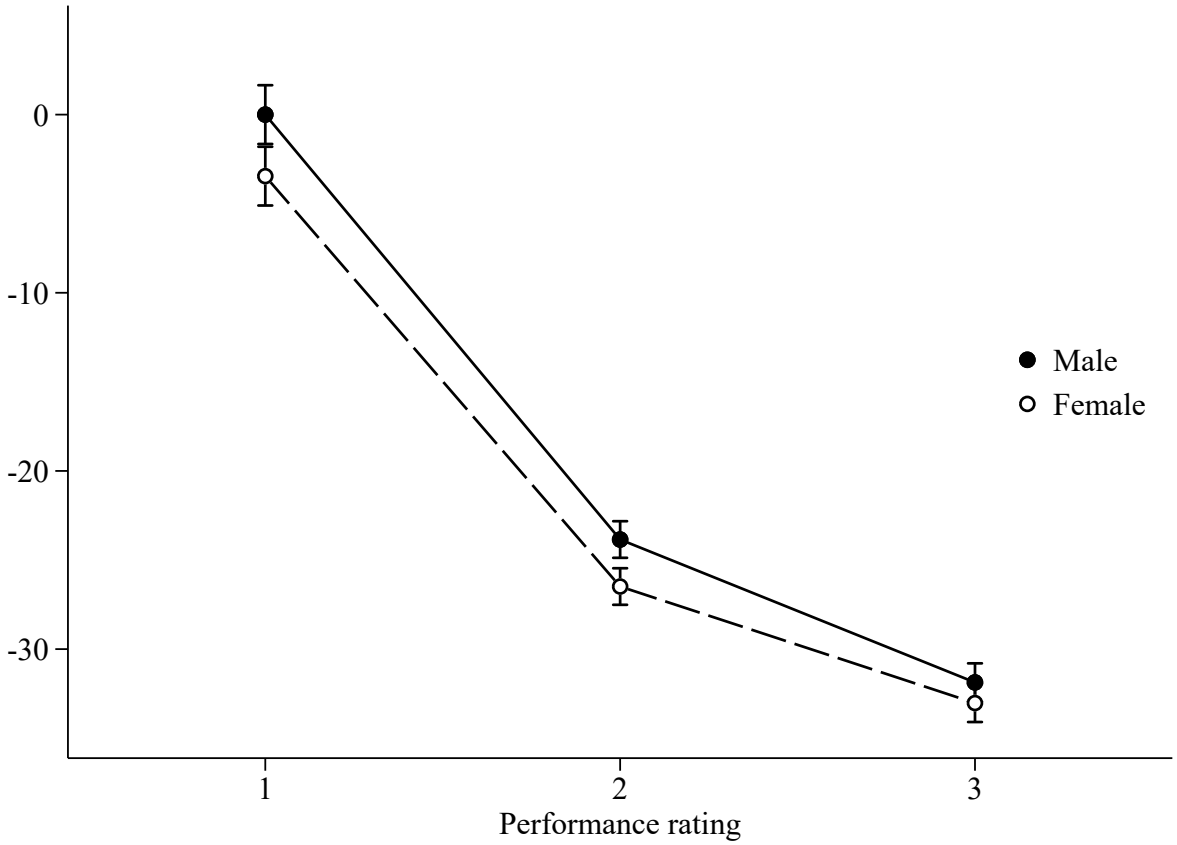
9 Appendix figures and tables

FIGURE A1: DISTRIBUTION OF NINE BOX RATINGS AND PROMOTIONS



NOTES: The top panel provides the share of men and women receiving each Nine Box rating. The bottom panel provides the annual promotion rate conditional on receiving each Nine Box rating for men and women. We exclude observations rated a low performance and high potential (the top left box) from our sample, because that rating is reserved by our firm for new hires.

Figure A2: Attrition by performance rating



Notes: This figure shows coefficient estimates from Table A6, which uses a linear probability model to estimate turnover by sex and performance rating. Point estimates are relative to men who receive as performance rating of 1; these men have an annualized attrition rate of 63.9%, versus 42.1% in the full sample, and 35.8% for high performing women (for whom attrition rates are lowest). Vertical brackets represent clustered standard errors.

Appendix Table A1: Decomposing the effect of ratings on the promotion gap

Panel A Interacted model	Coefficient	Standard error
Female	-1.382**	(.618)
Potential rating = 2	10.294***	(.375)
Potential rating = 3	18.306***	(1.038)
Female × Potential rating = 2	.508	(.595)
Female × Potential rating = 3	3.327*	(1.86)
Performance rating = 2	6.613***	(.435)
Performance rating =3	10.652***	(.56)
Female × Performance rating = 2	.626	(.658)
Female × Performance rating = 3	-1.238	(.822)
Panel B Decomposition	**Coefficient**	**Standard error**
Overall		
Men's promotion rate	12.623***	(.175)
Women's promotion rate	10.988***	(.196)
Gap	1.635***	(.263)
Gap explained by endowments		
Potential rating	.9***	(.07)
Performance rating	-.159***	(.022)
Gap explained by coefficients		
Potential rating	-.286	(.199)
Performance rating	-.144	(.606)

Notes: This table reports results from a Oaxaca-Blinder-Kitagawa decomposition. Panel A presents a pooled regression model for promotion where female is interacted with performance and potential, with fiscal year fixed effects. Panel B reports decomposition results. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Appendix Table A2: Variation by county characteristics

Potential rating	(1)	(2)	(3)
Female	0.135	0.124**	-0.296***
	(0.0937)	(0.0627)	(0.0521)
County management gap	-1.310***		
	(0.106)		
Female × County management gap	-0.379**		
	(0.159)		
County pay gap		-0.417***	
		(0.0303)	
Female × County pay gap		-0.151***	
		(0.0450)	
County female educational attainment			0.529***
			(0.0566)
Female × County female educational attainment			0.327***
			(0.0812)
Fiscal year FEs	Yes	Yes	Yes
Observations	780753	780753	780753

Notes: This table shows how the gender gap in potential ratings varies with county-level characteristics. *County management gap* is the fraction of men among workers with management standard occupational classification (SOC) codes. *County pay gap* is men's median earnings divided by women's median earnings. *County female educational* attainment is the fraction of women over the age of 18 with at least some college education. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Appendix Table A3: Over 50 sample

	(1) Potential rating	(2) Performance rating	(3) Log salary	(4) Promoted	(5) Promoted
Female	-0.0547***	0.0355***	-0.196***	-0.780**	-0.428
	(0.00721)	(0.00926)	(0.00928)	(0.321)	(0.312)
Potential rating					
2=Med					8.496***
					(0.502)
3=High					13.62***
					(2.116)
Performance rating					
2=Med					3.430***
					(0.458)
3=High					5.330***
					(0.550)
Fiscal year FEs	Yes	Yes	Yes	Yes	Yes
Observations	316890	316890	316445	316890	316890

Notes: This tables restricts the sample to workers aged 50 and over. Standard errors are clustered by worker. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Appendix Table A4: Promotions requiring relocation

Promotion	Same-state promotion		Out-of-state promotion	
	(1)	(2)	(3)	(4)
Female	-1.009***	-0.231	-0.828***	-0.731***
	(0.251)	(0.243)	(0.0885)	(0.0873)
Potential rating				
2=Med		9.404***		1.117***
		(0.277)		(0.0990)
3=High		17.19***		2.304***
		(0.826)		(0.321)
Performance rating				
2=Med	5.849***	6.179***	0.649***	0.677***
	(0.302)	(0.306)	(0.119)	(0.120)
3=High	10.16***	9.062***	1.182***	1.030***
	(0.395)	(0.388)	(0.150)	(0.149)
Fiscal year FEs	Yes	Yes	Yes	Yes
Observations	900209	900209	900209	900209
DV mean	10.619	10.619	1.33	1.33

Notes: Columns (1) and (2) indicate the worker is promoted. Columns (3) and (4) indicate the worker is both promoted and is working in the same state following the promotion. Columns (5) and (6) indicate the worker is promoted and working in a different state upon promotion. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Appendix Table A5: Performance in objectively measured sales positions

Sales	(1)	(2)	(3)	(4)
Female	0.0658***	0.0639***	0.0639***	0.0583***
	(0.00306)	(0.00304)	(0.00304)	(0.00303)
Fiscal year FEs	Yes			
Fiscal month FEs	Yes			
Location × month FEs		Yes	Yes	Yes
Job level FEs			Yes	Yes
Demographic controls				Yes
Observations	1844508	1844508	1844508	1844503
DV mean	1.124	1.124	1.124	1.124

Notes: This table uses OLS to compare mens' and women's sales performance. Sales performance is measured by the individual's hourly sales divided by the hourly sales target winsorized at 1%, where hourly goals are set by centrally based on factors such as location, department, and seasonality. Some specifications also include these as controls to allow for miscalibrated sales targets. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Appendix Table A6: Test of selection effects due to attrition of high performers

Attrition	(1)	(2)
Female	-3.452**	-2.080
	(1.649)	(1.638)
Potential rating		
2=Med	-1.329***	-2.810***
	(0.439)	(0.448)
3=High	0.286	-2.123**
	(0.926)	(0.942)
Performance rating		
2=Med	-23.85***	-21.52***
	(0.985)	(0.981)
3=High	-31.87***	-27.50***
	(1.019)	(1.027)
Female × Potential rating		
Female × 2=Med	1.220*	0.999
	(0.688)	(0.681)
Female × 3=High	2.048	2.005
	(1.545)	(1.538)
Female × Performance rating		
Female × 2=Med	0.815	0.158
	(1.667)	(1.649)
Female × 3=High	2.295	1.040
	(1.710)	(1.695)
Fiscal year FEs	Yes	Yes
Demographic controls		Yes
Location FEs		Yes
Observations	900209	900209
DV mean	42.125	42.122

Notes: This table presents results from a linear probability model where the dependent variable, attrition, takes values of 0 or 1,200 so that coefficients can be interpreted as annual percents. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Appendix Table A7: Robustness check using manager fixed effects

	(1) Performance rating	(2) Potential rating	(3) Promotion	(4) Next performance rating	(5) Next potential rating
Female	0.0388***	-0.0658***	-2.226***	0.0294***	-0.0508***
	(0.00509)	(0.00530)	(0.288)	(0.00505)	(0.00510)
Performance rating					
2=Med			6.352***	0.291***	0.196***
			(0.373)	(0.0115)	(0.00950)
3=High			10.92***	0.589***	0.307***
			(0.475)	(0.0125)	(0.0105)
Potential rating					
2=Med			9.543***	0.102***	0.323***
			(0.319)	(0.00513)	(0.00601)
3=High			18.08***	0.178***	0.581***
			(0.875)	(0.0117)	(0.0163)
Fiscal year FEs	Yes	Yes	Yes	Yes	Yes
Manager FEs	Yes	Yes	Yes	Yes	Yes
Observations	899581	899581	899581	586014	586014
DV mean	2.18	1.429	11.673	2.211	1.383

Notes: This table reproduces main results, but includes controls for the direct manager who provides the initial ratings. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Appendix Table A8: Robustness check using job level fixed effects

	(1) Potential rating	(2) Performance rating	(3) Promotion	(4) Next potential rating	(5) Next performance rating
Female	-0.0588***	0.0489***	-1.936***	-0.0428***	0.0346***
	(0.00548)	(0.00525)	(0.275)	(0.00481)	(0.00467)
Potential rating					
2=Med			8.714***	0.397***	0.0828***
			(0.295)	(0.00557)	(0.00491)
3=High			18.48***	0.695***	0.152***
			(0.840)	(0.0164)	(0.0117)
Performance rating					
2=Med			5.104***	0.189***	0.354***
			(0.341)	(0.00908)	(0.0112)
3=High			8.548***	0.287***	0.749***
			(0.426)	(0.00991)	(0.0121)
Fiscal year FEs	Yes	Yes	Yes	Yes	Yes
Job level FEs	Yes	Yes	Yes	Yes	Yes
Observations	900209	900209	900209	586338	586338
DV mean	1.429	2.18	11.949	1.383	2.21

Notes: This table reproduces main results, but includes controls for 22 job level fixed effects. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Appendix Table A9: Robustness check using pay decile fixed effects

	(1) Performance rating	(2) Potential rating	(3) Promotion	(4) Next performance rating	(5) Next potential rating
Female	0.0644***	-0.0556***	-1.467***	0.0466***	-0.0374***
	(0.00506)	(0.00550)	(0.263)	(0.00460)	(0.00477)
Performance rating					
2=Med			7.558***	0.346***	0.201***
			(0.345)	(0.0111)	(0.00914)
3=High			11.75***	0.730***	0.288***
			(0.443)	(0.0121)	(0.0100)
Potential rating					
2=Med			11.04***	0.0760***	0.412***
			(0.298)	(0.00487)	(0.00556)
3=High			20.42***	0.142***	0.709***
			(0.869)	(0.0115)	(0.0165)
Fiscal year FEs	Yes	Yes	Yes	Yes	Yes
Pay decile FEs	Yes	Yes	Yes	Yes	Yes
Observations	899023	899023	899023	585960	585960
DV mean	2.18	1.429	11.885	2.211	1.383

Notes: This table reproduces main results, but includes controls for 10 pay decile fixed effects. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Appendix Table A10: Robustness check using annual observations

	(1) Potential rating	(2) Performance rating	(3) Promotion	(4) Next potential rating	(5) Next performance rating
Female	-0.0813***	0.0355***	-0.923***	-0.0482***	0.0327***
	(0.00558)	(0.00520)	(0.216)	(0.00471)	(0.00458)
Potential rating					
2=Med			9.098***	0.424***	0.0910***
			(0.247)	(0.00551)	(0.00484)
3=High			16.89***	0.730***	0.167***
			(0.716)	(0.0167)	(0.0117)
Performance rating					
2=Med			6.036***	0.213***	0.363***
			(0.291)	(0.00916)	(0.0111)
3=High			8.488***	0.314***	0.766***
			(0.357)	(0.00997)	(0.0121)
Fiscal year FEs	Yes	Yes	Yes	Yes	Yes
Observations	79829	79829	79829	48920	48920

Notes: This table reproduces main results, but one observation represents one fiscal year rather than one month. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Appendix Table A11: Robustness check using manager-clustered errors

	(1) Potential rating	(2) Performance rating	(3) Promotion	(4) Next potential rating	(5) Next performance rating
Female	-0.0830***	0.0342***	-0.954***	-0.0482***	0.0327***
	(0.00530)	(0.00492)	(0.270)	(0.00450)	(0.00441)
Potential rating					
2=Med			10.51***	0.424***	0.0913***
			(0.326)	(0.00547)	(0.00460)
3=High			19.43***	0.730***	0.168***
			(0.913)	(0.0162)	(0.0107)
Performance rating					
2=Med			6.887***	0.213***	0.364***
			(0.353)	(0.00850)	(0.0102)
3=High			10.14***	0.315***	0.767***
			(0.450)	(0.00941)	(0.0115)
Fiscal year FEs	Yes	Yes	Yes	Yes	Yes
Observations	899581	899581	899581	586014	586014
DV mean	1.429	2.18	11.899	1.383	2.211

Notes: This table reproduces main results, but standard errors are clustered by manager who is rating the worker, rather than the worker themselves. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Appendix Table A12: Robustness check using combinations of potential and performance

	(1) Promotion	(2) Next performance rating	(3) Next potential rating
Female	-0.849***	0.0314***	-0.0454***
	(0.255)	(0.00454)	(0.00463)
Potential=1, Performance=2	4.805***	0.342***	0.0774***
	(0.351)	(0.0154)	(0.0107)
Potential=1, Performance=3	3.702***	0.802***	0.100***
	(0.426)	(0.0167)	(0.0117)
Potential=2, Performance=1	4.032***	0.0775***	0.132***
	(0.611)	(0.0210)	(0.0168)
Potential=2, Performance=2	13.26***	0.470***	0.469***
	(0.441)	(0.0158)	(0.0118)
Potential=2, Performance=3	21.66***	0.808***	0.665***
	(0.650)	(0.0169)	(0.0132)
Potential=3, Performance=2	19.82***	0.585***	0.731***
	(1.081)	(0.0210)	(0.0232)
Potential=3, Performance=3	31.52***	0.851***	0.974***
	(1.475)	(0.0234)	(0.0262)
Fiscal year FEs	Yes	Yes	Yes
Observations	900209	586338	586338

Notes: This table reproduces main results, but interacts potential and performance. The reference is Performance=1, Potential=1. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Appendix Table A13: OLS and IV estimates for future potential and performance

	Next potential		Next performance	
Model	Beta	S.E.	Beta	S.E.
a. Year and performance controls				
OLS female sample	.0312**	(.0123)	.0472***	(.0114)
OLS male sample	.0991***	(.0107)	.0384***	(.0089)
IV Female sample	-.0529**	(.0253)	.0337	(.0271)
IV Male sample	-.0231	(.0229)	-.0181	(.021)
b. Full controls				
OLS Female sample	.0313**	(.0124)	.0466***	(.0114)
OLS Male sample	.1003***	(.0107)	.0393***	(.0089)
IV Female sample	-.0587**	(.0263)	.0294	(.0275)
IV Male sample	-.0295	(.0237)	-.017	(.0212)

Notes: This table presents the coefficients on promotion for sixteen separate regressions described by equations (2) and (3). The regressions represent combinations of two outcomes (next potential and next performance), two models (OLS and 2SLS), subsamples for two sexes (female and male), and two sets of controls (fiscal year and past performance, then a full set of controls that adds demographics and location). The eight models for each women and men respectively have 228,680 and 315,104 observations. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Data appendix: County-level labor market gender inequality measures

We construct labor market gender inequality measures for US counties based on the methodology in Human Development Reports (2021). The county level variables were collected from the 2019 US Census Bureau five year estimates from the American Community Survey (2019). In Human Development Reports, gender-based inequality is measured using fifteen variables in three dimensions, including many measures focused on health, fertility, and mortality. We focus on three variables tied to labor market outcomes with a focus on upper level management: *County management gap* is the fraction of men among workers with management standard occupational classification (SOC) codes. *County pay gap* is men's median earnings divided by women's median earnings. *County female educational* attainment is the fraction of women over the age of 18 with at least some college education.