"Potential" and the Gender Promotion Gap*

Alan Benson
Univ. of Minnesota

Danielle Li
MIT & NBER

Kelly Shue
Yale & NBER

July 26, 2021

Abstract

We show that widely-used subjective assessments of employee "potential" contributes to gender gaps in promotion and pay. Using data on management-track employees from a large 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: women outperform male colleagues with the same potential ratings, both on average and conditional on promotion. Yet, even in these cases, women's subsequent potential ratings remain low, suggesting that firms persistently underestimate the potential of female employees.

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

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 Li for excellent research assistance and the International Center for Finance at the Yale School of Management for their support. We thank seminar audiences at INSEAD, Minnesota, SIOE, and Yale for helpful comments.

1 Introduction

When making promotion decisions, firms must make predictions about the future performance of each employee. If given the opportunity, would someone make a good manager? To guide their decisions, firms can use information about a worker's job performance. However, past job performance is an imperfect predictor of future performance both because people can grow over time and because management roles often rely on different skills compared to lower job roles (Peter and Hull, 1969; Benson, Li and Shue, 2019). Thus, firms need to make inferences about a worker's "potential." Unfortunately, firms can never directly observe a worker's potential and must instead make forecasts of potential that are highly subjective, leaving room for bias.

Promotion decisions that rely on subjective assessments of potential may negatively impact the careers of women through several related channels. Eagly and Karau (2002) finds broad evidence that qualities stereotypically held by women are incongruous with qualities stereotypically held by effective leaders, such as a results-orientation, assertiveness, ambition, competitiveness, creativity, and dominance (see also Bursztyn, Fujiwara and Pallais (2017); Koenig et al. (2011); and Proudfoot, Kay and Koval (2015)). Indeed, the simple fact that women are less frequently observed in managerial roles may make it more difficult to imagine women as having the skills for these positions (Blau and Kahn, 2017). It is also commonly alleged that "men are judged on their potential and women on their accomplishments"[1] and related research by Player et al. (2019) finds that experimental subjects assess greater leadership potential and forecast higher future performance for male applicants based on otherwise identical resumes. Finally, women may engage in less self promotion and networking (Cullen and Perez-Truglia, 2020; Fang and Huang, 2017), leading to a disadvantage in subjective assessments of potential that can be influenced by favouritism and politicking (Prendergast and Topel, 1993).

In this paper, we show that subjective assessments of worker potential contribute to gender gaps in promotion and 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 assessment and succession planning tool known as a "Nine Box" grid, in which supervisors rate subordinates on two dimensions: their current performance and their future potential. These dimensions take three values (low, medium, and high), creating a 3-by-3 matrix with nine cells, each attached to a label and a a prescription on how to lead them. Whereas performance ratings are intentionally backward looking and often based on demonstrable achievements, potential ratings evaluate a worker's capacity for future growth and development, making them fundamentally more subjective. Beyond their use at our firm, Nine Box and similar assessments of workers' potential

[1] Goldberg, Michelle. "There Could Never Be a Female Andrew Yang." *New York Times* 19 April 2021.

are nearly ubiquitous in large organizations, where they often play a major role in determining promotions, developmental opportunities, and compensation.[2]

Our paper has three key findings. First, women are significantly less likely to be promoted despite receiving higher performance evaluations. Second, women are rated as having lower potential both on average and conditional on their performance. We estimate that the difference in rated potential explains up to half of the overall gender gap in promotions, with the remainder coming from the direct effect of being female and correlated demographics. Lastly, women's lower average potential ratings do not appear to reflect lower average future performance; among workers with the same current performance and potential ratings, women receive better performance ratings in the future. Likewise, promoted women outperform promoted men, consistent with firms holding women to a higher standard during the promotions process. Moreover, women continue to receive lower potential ratings after promotion, implying that the biased evaluation of potential persists even when women have outperformed prior forecasts.

We motivate our analysis by documenting 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 intensify 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 is consistent with 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 these aggregate differences, we find a robust gender gap in promotions at the individual level: women are 14% less likely to be promoted relative to the sample average. This baseline gender gap in promotions cannot be explained by differences in past performance: women receive higher performance ratings on average and are 7.3% more likely to earn the top performance rating relative to the sample average.

We show that the gender gap in promotions is better explained by differences in forecasts of potential. At our firm, potential ratings strongly predict promotions, even more so than performance ratings. An increase from "medium" to "high" potential ratings corresponds to a 75% increase in the likelihood of promotion relative to the base rate, compared to only a 27% increase for the equivalent increase in performance ratings. Women in our firm receive 8.3% lower potential ratings than men on average, despite their higher performance ratings, and despite the fact that performance and

[2] Cappelli and Keller (2014) and Church et al. (2015) discuss the history of Nine Box, its widespread adoption, and other assessments of employee potential.

potential ratings are highly positively correlated within individuals. Taken together, we find that gender differences in potential ratings can explain up to 50% of the overall gender promotion gap.

Women's lower potential ratings, however, may be justified if they do in fact 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 scores in their next evaluation. This difference in future performance is consistent with a model of discrimination in which women are held to a higher standard: to receive the same potential score as their male colleagues, women's future performance must be higher. This result holds whether we consider worker's future performance in the same role or whether we consider the performance of men and women who are promoted into higher-level managerial roles.

In addition, we find that women continue to receive lower potential scores even after demonstrating stronger future performance. That is, when women outperform men with identical past potential and performance ratings, women continue to receive lower potential ratings in the current period. This result runs counter to the idea that women's lower potential ratings are driven purely by a lack of data on women in high-level roles. In our data, when promoted women receive higher performance ratings in their new roles (relative to promoted male colleagues with the same pre-promotion scores) these women still receive lower potential scores going forward. Indeed, we show that the gender gap in potential ratings actually grows as one advances in the organizational hierarchy. This fact accelerates the already widening gender gap that comes from the simple compounding of a constant gender gap in promotions as one advances up the career ladder.

Throughout our analysis, we distinguish between two types of gaps: the unconditional gap in promotion, ratings, and performance between men and women; and the gap conditional on factors such as other demographics and retail store assignment. We find that the raw female disadvantage in potential ratings and promotion is partly explained by the fact that women in our sample are more likely to be older, have longer tenure within the firm, and to belong to minority ethnic groups, all of which are factors associated with lower potential ratings and promotion rates. However, the gender gap remains large after controlling for these other demographic variables, and we do not observe significant interaction effects between gender and other demographic characteristics (the "double jeopardy" hypothesis). We also find that differences in store assignment by gender appear to be small and do not account for large differences in potential or performance ratings.

Next, we examine heterogeneity in ratings, pay, and promotion gaps. We highlight two particular dimensions of heterogeneity—manager gender and manager performance—because it it 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 supportive of their female subordinates and high performing managers may be better able to evaluate their subordinates in an unbiased

manner. Our data contain information on the demographics and own Nine Box ratings for each worker's manager, allowing us to evaluate how manager characteristics correlate with their subordinate job outcomes. In these analyses, we caution that we cannot distinguish between treatment and selection effects; that is, subordinate outcomes may differ across managers because managers are assigned to different types of subordinates or because managers differ in how they evaluate their subordinates.

We find that gender gaps in potential, pay, and promotions are only slightly smaller under female managers. Moreover, female managers are associated with lower overall potential ratings, pay, and promotion rates for their subordinates of both genders. Conversely, we find that gender gaps in potential, pay, and promotions are larger under more highly rated managers, but 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 mean that it is not obvious whether female subordinates would be better off working under either female managers or highly rated managers.

In the final section of our paper, we consider the impact of counterfactual policies in which we vary the way that potential ratings are assigned and used in promotion decisions. We first consider what happens if firms make promotion decisions on the basis of performance ratings alone, ignoring information on potential and gender. We show that this policy would nearly eliminate the gender promotion gap, but also decrease the average future performance of workers who are selected to be promoted. The latter result reflects the fact that, despite their bias, potential ratings contain useful information about future performance. Ignoring potential ratings would reduce gender bias, but also reduce the quality of the firm's information.

Next, we consider policies that retain information on potential, but simply increase the potential ratings of women by one point (conditional on the current rating being below the maximum). We can apply this score correction to all women or only to women with the highest performance rating. We find that the latter approach equalizes promotion rates by gender while also increasing the predicted future performance of promoted workers. While this simple policy may be challenging to implement in practice (for instance, managers may shade female potential ratings down in anticipation of this gender-specific bonus), it suggests that firms may be able to increase the quality of their promotion decisions by finding ways to de-bias their otherwise informative potential scores.

Taken together, our results show that firms face various challenges related to the objectivity, informativeness, and equity of their workplace evaluations. Promotions based solely on performance data may be more objective but are not the most informative. This is consistent with research on the "Peter Principle," which has shown that performance in one's current role is a highly imperfect predictor of one's potential to succeed as a manager (e.g., Benson, Li and Shue, 2019). To identify the best managers, firms need to find alternative ways of forecasting a worker's management potential.

Yet, solutions such as the Nine Box system that have arisen to address the Peter Principle problem rely on subjective assessments of potential which we show to be biased against women. Reducing bias in assessments of potential could improve both the equity and efficacy of promotion decisions. At the very least, we argue that subjective assessments of potential need to be more carefully scrutinized before being codified into tools such as the Nine Box grid, which may lend a false sense of legitimacy to biased ratings and widen the gender promotion gap.

2 Background

2.1 Setting

Our data come from the U.S. operations of a large retailer from February 2009 to October 2015. Over this period, the firm employed over one million workers, nearly one percent of the US labor force, 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, nearly all of whom are evaluated using Nine Box ratings. Our main sample, therefore, consists of 29,809 management-track workers, 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 in information technology, supply chain management, finance, human resources, and real estate management. Career ladders follow a traditional system of pay grades nested within bands. 40% of corporate workers with Nine Box scores are categorized as individual contributors, 40% are managers, and 20% 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 work at one of over 4,000 establishments. Most employees work in a store that is led by a head store manager and a team of department managers. Store managers report to a senior manager who covers all stores of a given format in one of the country's 37 districts. Store managers are primarily responsible for leadership and management activities, including analyzing data, formulating a strategy, and inspiring others to successfully execute that strategy. Store managers are assessed on their ability to achieve performance goals but are otherwise given wide latitude in how to achieve them. Department managers are responsible for efficiently executing the strategy set out by the store manager. This includes customer-facing duties and supervisory duties such as the hiring and coaching of entry-level staff.

2.2 Nine Box evaluation

Employees in our data are evaluated using a Nine Box grid, a widespread talent assessment tool that instructs supervisors to categorize their subordinates into one of nine boxes representing the

interaction of three categories (low, medium, and high)for the worker's prior job performance and three categories (low, medium, and high) for the worker's future potential. Nine Box ratings are typically used for succession planning and the allocation of training, development, and promotion opportunities, but can also be tied to compensation.[3] 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 technical fields like science, engineering, law, and academia where the skills required to perform and manage a job are quite different (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 argue that Nine Box is less transparent, objective, impartial, 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. For this reason, traditional assessments of potential have relied on personality or IQ. More recently, 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 Nine Box potential ratings to be highly subjective.

Indeed, Nine Box's reliance on disaggregated, non-expert supervisors who are provided limited guidance, weak criteria, and little or no evidence makes it prone to well-documented biases. Most practitioner literature refers to performance as being a backward-looking measure, instructing raters to assess how the worker has performed in the current role. In contrast, the potential rating is nominally the forward-looking measure and instructs raters to assess the capacity for growth within the same role or in the organization, though these criteria themselves are often fluid even within organizations (Silzer and Church, 2009). In these circumstances, raters may refer to prior years' potential ratings (the anchoring bias), first impressions (the primacy bias), last impressions (recency bias), and nonrated criteria (halo bias) (for a review, see Kahneman, 2011).

Moreover, academic research points to several channels through which biases in subjective evaluations may work against women. First, the psychology literature on role congruity theory,

[3]For instance, Microsoft has also used performance ratings to distribute cash bonuses and potential ratings to allocate promotions (Bartlett, 2001).

first proposed by Eagly and Karau (2002), has argued that people may have a hard time imagining women as qualified for leadership positions because of the mismatch between traditional female stereotypes and traditional leadership stereotypes. Second, subjective evaluations may be influenced by politicking and favoritism (Prendergast and Topel, 1993). This could translate to a disadvantage for women if they have less access to networking opportunities (Cullen and Perez-Truglia, 2019), benefit less from connections (Fang and Huang, 2017), or engage in less self promotion.[4] Finally, self-interested managers may manipulate potential scores to keep their best subordinates. 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.

Despite these concerns, Nine Box remains a highly popular method of identifying candidates for developmental opportunities and promotion. 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."[5]

At our firm, Nine Box ratings are assigned annually in a two-step process. First, managers provide initial ratings for all of their salaried subordinates. After this, there are a series of district- and headquarter-level calibration meetings, during which scores may be adjusted to ensure that similar standards are applied to employees doing similar types of work. Although we are not aware of any systematic surveys that examine exactly how organizations use Nine Box, we believe our firm is typical in its two dimensional rating, the labels it ascribes to each of the nine boxes, and the reliance on immediate supervisors rather than objective data or psychological tests for initial ratings. Upper-level Nine box calibration meetings also appear to be commonplace at large organizations; such measures suggest that our firm provides similar or greater oversight than the completely disaggregated Nine Box procedures sometimes described in the practitioner literature.

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 employees at our sample firm during this period. Our data includes

[4] Azmat, Cuñat and Henry (2020) find that female lawyers report lower partnership aspirations. In our setting, low aspirations may translate into women engaging in less self promotion, which could, in turn, negatively impact their potential ratings. We note that possible gender differences in aspirations are compatible with a view in which promotions are biased against women. Azmat, Cuñat and Henry (2020) model aspirations as endogenously determined by workplace gender discrimination. Further, we find in our retail setting that women receive higher performance ratings than men, suggesting that any gender differences in aspirations do not translate into lower effort provision by women.

[5] We are not aware of any systematic studies of Nine Box's adoption. Practitioners we spoke to at CitiGroup, Bristol Myers Squibb, Honeywell, 3M, Ecolab, and General Mills confirmed its widely known and used, including at their organizations.

workers employed in the firm's corporate headquarters, as well as workers employed across 4,101 retail locations.

Our main data is at the worker-month level. We code a worker as being promoted if we observe an upward change in job levels in the next month. Examples of promotions in our data include moving from department manager to store manager, or moving from web developer to lead web developer.

Nine box assessments are finalized and recorded in the fourth quarter of each fiscal year, which ends in January. We set workers' Nine Box ratings in each month equal to her rating for the corresponding fiscal year. That is, if Nine Box scores are assigned in Q4 of 2019, in reference to the past fiscal year's activities, then a worker's May 2019 rating will be set equal to the ratings she receives in Q4 of 2019.

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

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 received a promotion if she experiences a change in job title that is associated with an increase in average compensation associated with that job title or experiences a change in job title that is associated with a personal raise in salary exceeding 5%.

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. As a simple preview of our more detailed empirical results, it is evident from this panel that being female is positively correlated with performance ratings and negatively correlated with promotion, annual salary, and potential ratings.

3 Potential and the Gender Promotion Gap

In this section, we present several results aimed at documenting how potential ratings can help explain the gender promotion gap. We begin by describing promotion rates in our sample firm, both

in the raw data and controlling for various worker characteristics. Next, we document the gender gap in potential and show that it can explain a substantial portion of the overall gender promotion gap.

3.1 Gender and Promotion

We begin by documenting differences in gender representation. In our firm, as in many others, the share of women progressively decreases as one ascends the career ladder, as illustrated in Figure 1. In the left panel, we focus on workers in retail operations, for which there exists a clear ordering of job titles. Here, women make up 56% of entry level workers (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.

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 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 score, age and race, 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 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, -1.6, on the female indicator implies that women are 1.6 percentage points less likely to be promoted, or 13.5% less likely to be promoted related to the base promotion rate 11.9%. Because this difference in promotion could be due to differences in performance, we control for the worker's Nine Box performance ratings fixed effects 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 larger female disadvantage in promotions.

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 B, women tend to be older and have longer tenure within the firm, and to be Black or Hispanic; 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).

Women may also be assigned to different store and administrative locations than men. In Column 4, we control for location fixed effects, and find the gender promotion gap remains approximately constant in size if we compare workers in the same location. To offer a comprehensive view of the data, we present both the "overall gender promotion gap," which refers to the raw gap in Column 1 and the "gender promotion gap, controlling for performance, demographics, and location," which refers to the estimates in Column 4, after including performance, demographic, and store location controls. We will continue to estimate specifications that include these control variables throughout our analysis.

Table 3 documents how differences in promotion rates can 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. While these variable do significantly predict compensation, they do not lead to large changes in our estimate of the gender wage gap. Instead, gender 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 why women have lower promotion rates. Table 4 considers how Nine Box performance and potential ratings differ for men and women. Panel A measures ratings on a 1, 2, and 3 scale while Panel B looks at differences in the probability of earning the top rating of 3. We find that women receive substantially higher performance ratings, both in the raw data and conditional on demographics and location. In particular, Column 1 of Panel B shows that women are 1.81 percentage points more likely to earn the top performance rating, an increase of 7% relative to the base probability of earning the top performance rating.

In contrast, women earn substantially lower potential ratings, both in the raw data and conditional on demographics and location. Column 3 of Panel B shows that women are 1.43 percentage points less likely to earn the top potential rating. This gap is quite large compared to the base probability of earning the top potential rating (4.46%)—women are 32% less likely to earn the top potential rating. The divergence in potential and performance ratings for women is all the more surprising because that the two ratings are positively correlated in the overall sample, as shown in Table 1 Panel B.[6] This divergence suggests that potential scores may be biased against women, a question we evaluate in more detail in Section 4.

Figure A1 plots additional details about the gender difference in performance and potential scores. The left panel plots the distribution of performance and potential scores for men in our sample, while the right panel represents the differences in shares for women relative to men. Women are significantly less likely to earn low performance ratings and significantly more likely to earn high performance ratings. The opposite pattern occurs for potential ratings. Women are significantly more likely to earn the lowest potential rating and significantly less likely to earn the highest potential rating.

In Table 5, we examine the extent to which the gender gap in potential ratings explains the gender gap in promotion. We replicate each column of Table 2, adding controls for the 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. 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.

[6] 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.

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 scores, demographics, and location assignment. Second, Table 5 shows that potential ratings are very strong predictors of promotion; in fact they matter more than performance ratings. 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 promotion rate, while a similar change in performance ratings from 2 to 3 corresponds to only a 3.24 percentage point increase in the promotion rate.

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 push for or accept promotions. These gender differences in career aspirations may arise endogenously from other forms of 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 is likely endogenous. Nine Box potential ratings at our firm (and by convention) are used to allocate scarce internal developmental opportunities, and the prospect of being high-performing but "invisible" can reduce incentives for women and other minorities to invest in development themselves Milgrom and Oster (1987).

4 Informativeness of Potential Assessments

So far, we have shown that low assessments of potential help explain why women are less likely to be promoted. In this section, we examine the information content of potential assessments and show that, despite containing useful information about a worker's future performance, potential scores appear to be "biased" against women.

4.1 Potential scores and realized future performance

We begin by assessing the predictive value of potential scores. 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 future performance, particularly among the sample of workers who are promoted into management positions.

Table 6 Panel A shows that high potential ratings in the current year predict higher performance scores in the following fiscal year, even after conditioning on the worker's current performance score. This positive relation holds both in the full sample of workers, as well as within the subsample of employees who experience a promotion event. Thus, potential ratings appear to contain real information regarding the worker's future performance in general, and conditional on promotion into a higher-level position.

4.2 Gender bias in potential scores

If potential scores are predictive of future performance, then one natural explanation for why women receive lower potential scores is that they indeed have lower future performance.[7] If potential scores are justified, then men and women with the same current period potential scores should have similar future performance ratings. Following the logic of a Becker outcomes test for discrimination, assessments of potential are biased against women if, for the same potential score, women have higher average future performance ratings.

Table 6 Panel A examines this possibility by relating current period potential scores with next period performance. 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 scores, 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 fiscal year 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 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. This finding is consistent with discrimination in the promotion

[7] 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 women Uber drivers earn less per hour than men despite identical pay contracts due to differences in experience and driving preferences. However, Sarsons (2017) finds female surgeons are slightly higher ability than male surgeons.

decision: if promotions discriminate against women, then promoted women should exhibit superior performance as managers, which is exactly what we observe in the data.

Next, we consider how firms update their evaluations of potential in response to realized future performance. To do this, Table 6 B replicates the analysis in Panel A with potential scores in the next fiscal year as the outcome of interest. We find that women continue to receive significantly lower future potential ratings compared to men with identical current performance and potential ratings, both in the full sample and in the sample of newly promoted workers. Because performance and potential together comprise the firm's Nine Box evaluation, we know that performance and potential ratings are determined during the same meetings, by the same set of managers. This means that, at the same time that women are given performance ratings indicating that they outperformed their previous year's potential scores (relative to men with the same potential scores), women are still assessed as having lower potential going forward. The evidence in Table 6 suggests that gender gaps in performance and potential evaluations persist after each round of promotion decisions, and help to explain the decreasing representation of women while climbing each rung of the career ladder.

5 Additional results

5.1 Retention and Leave of Absences

One possible explanation for our results is that potential ratings reflect a manager's expectation of a worker's commitment to remain at the firm. If managers expect that women's careers are more likely to be interrupted by family care duties, then this may impact their assessments of potential, independent of women's current performance.[8] Our results in Section 4 focus on the relation between potential evaluations and future realized performance among employees who remain at the firm, but does not account for the possibility of differential attrition or leaves of absence. In this section, we explore whether the gender potential gap that we document can be explained by differences in these additional factors.

In Columns 1 and 2 of Table 7, we begin by showing that women are more likely to take future leaves of absence than their male colleagues, but less likely to leave the firm altogether. In Columns 3-5, we explore how these impact the gender potential gap. In particular, the gender gap in potential ratings appears unaffected by a worker's leave history, or even their realizations of leaves or attrition in the future. Column 4 shows that women continue to receive lower potential ratings, even after controlling for their past history of leaves or their future attachment to the firm. Column 5 shows

[8] We leave aside the important question of whether it is legal or ethical to consider future leave, particularly maternity leave, when forming Nine Box ratings or promotions decisions. In this paper, we focus on the empirical question of whether women in our sample take substantially more leave or are be more likely to exit the firm and whether that information appears to be correlated with potential ratings.

that these same results hold after restricting the sample to workers over the age of 45 who have not previously taken any leave; if firms were using potential ratings to make inferences about issues related to maternity leave, then we would not expect to find a gender potential gap in this sample.

5.2 Widening gender gaps over the career ladder

Our results so far indicate that gender gaps in potential ratings help explain gender gaps in promotions and do not appear to update in response to realized high future performance. If gender gaps in potential and promotion are constant or grow as one advances up the career ladder, a compounding effect would generate wider gender gaps in representation at top levels, consistent with what we observe in Figure 1. On the other hand, one might expect gender gaps in potential ratings and promotion to close as one advances toward corporate headquarters, because of the growing popularity of corporate diversity programs that aim to support female leaders.

To examine variation over the career ladder, Table 8 explores how the gender gap in performance, potential ratings, pay, and promotion rates changes as one advances in the organizational hierarchy. Specifically, we regress workers' performance and potential ratings, as well as salary and promotion outcomes, on the interaction of a worker's gender and their career ladder position as proxied by the worker's decile in the firm's pay distribution in that year, controlling for the individual effects of these variables. In all specifications, we control for fiscal year, demographics, and location fixed effects.

Our results indicate that, while women continue to outperform men throughout the hierarchy (and this gap does not change significantly), the gender gap in potential ratings widen at higher levels of the firm's hierarchy. This widening gender gap in potential occurs alongside a widening gender gap in promotions, which holds even after controlling for performance ratings.We also find that women continue to be paid less than their male colleagues, although this pay gap stays constant with respect to pay decile.

We note that a simple compounding effect implies that, given a constant gender gap in promotions, female representation will decline with each increasing rung in the career ladder. The fact that the gender gap in potential ratings and promotions grows with pay decile implies that the gap in female representation not only grows, but actually accelerate. Taken together, these results show that women face growing disadvantages as they advance up the career ladder.

5.3 Heterogeneity by manager assignment

In this section, we consider 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 performance and potential ratings. Our 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.[9] Likewise, women may benefit from working under higher quality managers who may be better at assessing their subordinates' true performance and potential, less concerned about competition from their own subordinates, and 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 performance and potential rating), and the interaction between worker gender and manager characteristics. We control for year, store location, other worker demographics and, in some cases, worker Nine Box ratings.

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 or because managers differ in how they assess or advocate for their subordinates.

In Table 9, we examine whether a subordinate's rating depends on their 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). In Column 1, we find that female workers earn higher performance ratings than their male colleagues, and this outperformance does not vary with the manager's gender. In the other columns, we find that gender gaps in potential ratings, pay, and promotions are slightly smaller under female managers. However, female managers are also associated with lower overall levels of ratings, pay, and promotion rates for all their subordinates, regardless of subordinate gender. This is evidenced by the negative coefficients on the manager female gender indicator. Taken together, the opposing level and interaction effects imply that it is not obvious that female employees would be better off working under a female manager. The female employee can expect a smaller gender *gap*, but not necessarily an increase in the absolute levels of ratings, pay, or promotion rates. Our findings of offsetting effects echoes related results in Cardoso and Winter-Ebmer (2010), who show that female-led firms are associated with lower gender wage gaps as well as levels of wages.

In Table 10, we explore how worker outcomes vary with their manager's performance and potential evaluations. We find that subordinates assigned to managers with higher performance and potential ratings are significantly more likely to receive higher performance and potential ratings, have higher salaries, and to be promoted. To the extent that these differences represent treatment effects, female workers would benefit from being assigned to more highly rated managers. We find,

[9]Research on the "queen bee" syndrome shows that female managers can sometimes be tougher on their female subordinates, possibly due to a competition effect (see e.g., Ellemers et al. (2004)). Thus, it is not obvious that female subordinates would be better off worker under a female manager.

however, that the interaction effect between worker gender and manager ratings are negative in almost all cases. This implies that although the level of worker outcomes is higher, gender gaps among subordinates assigned to more highly rated managers are also larger. On net, it is unclear whether women benefit from these assignments.

6 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 future performance of promoted candidates.

We consider the following counterfactual policies. First, given that we have documented evidence of gender bias in potential scores, one potential remedy is to simply stop using them in promotion decisions. The first counterfactual promotion policy we consider is to remove information on gender and potential from promotion decisions. Another possibility is to continue using potential scores, but to first "adjust" them to account for gender bias. We consider two very simple ways of accomplishing this task: the first is to add one point to the potential scores of female candidates, so that women who are rated "low" are now rated "medium," and those rated "medium" are now rated "high" potential. Since the maximum possible potential score is "high," we leave the potential scores of women who receive "high" potential scores unchanged. The second approach is a milder version of this correction, which only adds one to the potential ratings of women who are rated "high" in terms of performance.

To assess the impact of these promotion policies, we begin by estimating a regression of promotion on the female indicator, potential ratings dummies, performance ratings dummies, demographics, and year fixed effects. The coefficients from this regression represent the firm's actual promotion policy given a set of ratings and other worker characteristics. We then predict a worker's likelihood of promotion using the fitted value from this regression given the new inputs used by the proposed counterfactual policy.

That is, to evaluate the impact of 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 policies in which we adjust potential ratings, we simply use our counterfactual adjusted potential scores as inputs into predicting promotion likelihood.

Once we have fitted measures of a worker's likelihood of promotion under each policy, we assess the impact of each counterfactual policy on the expected promotion rates of men and women, as well as on estimates of the average future performance of promoted candidates. To form the average future performance of promoted candidates under each counterfactual policy, we use the weighted

average of workers' next period performance ratings, where the weights are the worker's 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. We can compute expected next performance ratings using the weighted average over either the full sample or the sample of true promotions. The full sample offers more complete coverage, while restricting the sample to true promotion events may offer a better measure of how future performance changes after a promotion event.

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 and 4 present the average future performance rating among promoted workers under each promotion policy. Focusing on the first two rows, we show that blinding promotion decisions to gender and potential ratings does indeed reduce the gender gap in promotions by 65%, from a 1.7 percentage point annualized gap to a 0.6 percentage point gap. The gender gap does not disappear because we do not blind the firm to other demographics that are correlated with gender and associated with promotion rates. Yet, we also find that this reduction in the gender gap would come at the expense of reducing the expected average future performance of workers who are more likely to be promoted. Specifically, in Columns 3, 4, and 5, we report weighted averages of next period performance over the full sample of workers, the sample of workers who were not promoted, and the sample of workers who were promoted. respectively. In all cases, we find that expected next performance ratings decline relative to the baseline promotion policy shown in the top row.

These results are consistent with Table 6, which shows that, despite being biased, potential ratings do contain useful information about future perfor mance. A promotion policy that ignores potential scores altogether would be based on a coarser information set that one that is able to incorporate the information in potential scores in some way.

Next, we consider what happens when we retain information on potential scores, but apply adjustments aimed at "undoing" gender bias. We show that uniformly increasing the potential scores of all women leads to a substantial reversal of the gender promotion gap, so that now women are 40% more likely to be promoted than men. This approach, however, also leads to a small decrease in the expected quality of workers who are more likely to be promoted. Finally, in our third counterfactual, we show that a more targeted shift in potential scores, applying only to women who are rated highest in terms of performance, leads to an improvement in both equity and efficiency. In particular, this approach eliminates the gender promotion gap, while also increasing the estimated next period performance of promoted workers.

We acknowledge that this analysis is subject to two limitations. First, while we observe next period performance for most workers regardless of promotion, next period performance holding the

job role constant may differ from next period performance conditional on promotion. Therefore, we use the full sample in Column 3 and the true promoted sample in Column 4 to create weighted averages for the expected next performance rating. Our results in Column 4 are based only on the subset of workers who the firm saw fit to promote, and thus may offer a more realistic measure of how performance may change after promotion. However, if women are positively selected into the true promoted sample relative to men (e.g., due to discrimination against women in the promotion process), then a counterfactual promotion policy that increases the weights on women within the true promoted sample may overstate the efficiency gains of the counterfactual policy when applied to the entire sample.

Second, the "adjustment" policies that we evaluate may be circumvented if managers lower scores for female subordinates in anticipation of them receiving a gender-specific bonus. Given this, we regard our third counterfactual not 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 de-bias assessments of potential, rather than getting rid of potential assessments entirely.

7 Conclusion

In this paper, we provide evidence that subjective assessments of worker potential, widely used for career planning within firms, contribute to persistent gender gaps in promotion and pay. Despite being more likely to receive top performance ratings, women are less likely to be thought of as having high "potential." These lower potential ratings can explain as much as 50% of the observed gender gap in promotions.

Women's lower potential ratings may be justified if they accurately forecast worse performance in the future. We show that this is not the case. Rather, we find that potential assessments, while informative about future performance on net, are biased against women. That is, among employees with the same current performance and potential ratings, women outperform men on evaluations of their future performance. This is consistent with classic discrimination models in which, to receive the same potential rating, women are held to a higher bar in terms of their expected future performance. We show, further, that this bias in potential ratings does not appear to be self correcting: even though women outperform their potential ratings, they continue to receive lower potential evaluations in the future. This persistence of lower potential ratings is true for women who continue in their current roles, as well as for women who are promoted and perform well in their new role.

Taken together, our results suggest that subjective assessments of potential are an ever present barrier to women's advancement in their careers. In our data, the gender gap in promotions widens

as one moves up the career ladder, as does the gender gap in potential scores. The failure of firms to update potential ratings to be in line with realized future performance suggests that inaccurate stereotypes and other types of biases may limit firms' abilities to accurately forecast actual potential.

We also find that biased evaluations of potential are challenging to address. First, one cannot simply decrease the gender promotions 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 scores to women conditional on performance. This suggests that policies that seek to decrease the gap between assessed potential and future performance need to address broader organizational questions, rather than simply changing the gender of the evaluator.

Similarly, we also show that assigning women to higher quality managers would not reduce gender bias. While managers who themselves receive higher performance and potential scores 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 results show that firms should not simply do away with potential ratings altogether. A growing 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, assessments of potential are predictive of future performance beyond what can be predicted by current performance. This means that potential scores add value despite their gender bias.

Instead, our results show that there may be large gains from finding ways to de-bias assessments of potential, for instance by reducing reliance on 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.

References

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.*

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.

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

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.

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.

Cullen, Zoë B, and Ricardo Perez-Truglia. 2019. "The Old Boys' Club: Schmoozing and the Gender Gap." National Bureau of Economic Research.

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

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.

Ellemers, Naomi, Henriette Van den Heuvel, Dick De Gilder, Anne Maass, and Alessandra Bonvini. 2004. "The underrepresentation of women in science: Differential commitment or the queen bee syndrome?" *British Journal of Social Psychology*, 43(3): 315–338.

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

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.*

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

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.

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

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.

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

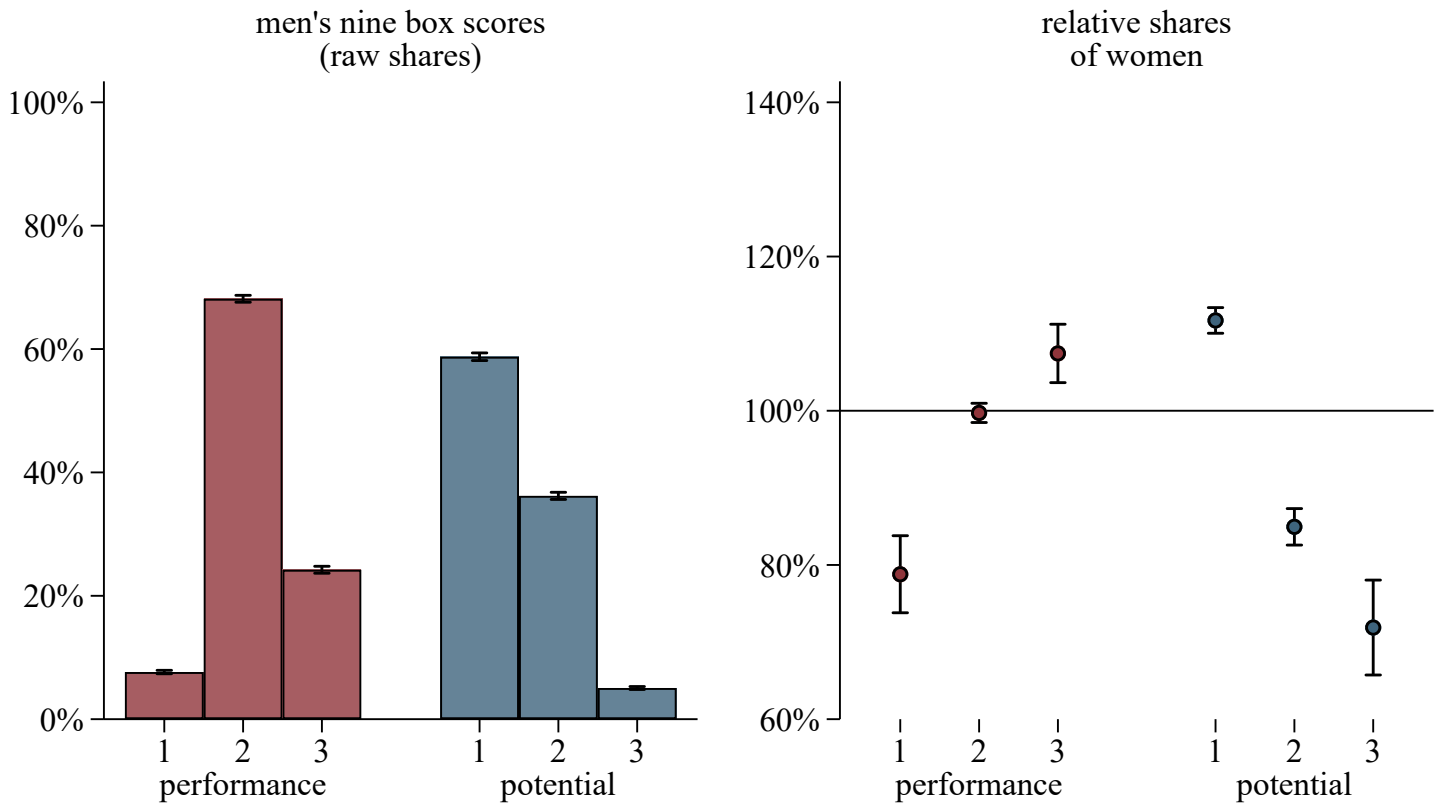
Silzer, Rob, and Allan H Church. 2009. "The pearls and perils of identifying potential." *Industrial and Organizational Psychology*, 2(4): 377–412.

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

FIGURE 1: FEMALE SHARES IN THE ORGANIZATIONAL HIERARCHY

NOTES: The left panel reports 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 that level. 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 2: GENDER GAP IN NINE BOX SCORES



NOTES: The left panel represents the distribution of Nine Box performance and ratings assigned to male employees. The right panel represents the raw difference in shares for female employees. Bars represent 95% confidence intervals.

Table 1: Descriptive statistics

Panel A: Data coverage						
Locations	$> 4,000$		Worker-months			900,209
Workers	29,809		Promotions			8,964
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)	Potential rating	1.429	.578	1	1	2
(f)	Age	44.4	10.834	35.8	45	53

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)	Potential rating	-.071	.048	.176	.088	1	
(f)*	Age	.037	-.052	.193	.015	-.271	1
(g)*	Tenure (months)	.072	-.039	-.021	.097	-.215	.465

Panel A reports data coverage. Tables and figures are estimated using the baseline sample consisting of observations at the worker-month level unless otherwise noted. Panel B presents summary statistics of variables used in our analysis. Panel C presents correlations between these variables. 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)
Log tenure			-2.377***	-1.825***
			(0.142)	(0.164)
Demographic FEs	Yes	Yes	Yes	Yes
Fiscal year FEs	Yes	Yes	Yes	Yes
Location FEs				Yes
Observations	900209	900209	900209	900209

This table reports regressions of an indicator for whether the worker is promoted on the female indicator, performance rating indicators (the omitted category is 1=Low), and other control variables for worker demographics, fiscal year fixed effects, and location fixed effects. 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***	-0.0376***	-0.0416***
	(0.00566)	(0.00391)	(0.00382)	(0.00367)
Performance rating				
2=Med			0.0537***	0.0653***
			(0.00432)	(0.00422)
3=High			0.149***	0.153***
			(0.00525)	(0.00508)
Potential rating				
2=Med			0.0468***	0.0829***
			(0.00294)	(0.00287)
3=High			0.0782***	0.136***
			(0.00654)	(0.00633)
Log tenure				-0.0192***
				(0.00183)
Fiscal year FEs	Yes	Yes	Yes	Yes
Demographic FEs	Yes	Yes	Yes	Yes
Job level × year FEs		Yes	Yes	Yes
Observations	899023	899023	899023	899023

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), and other control variables for worker demographics, fiscal year fixed effects, and location fixed effects. 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	Performance rating		Potential rating	
	(1)	(2)	(3)	(4)
Female	.0343***	.0151***	-.083***	-.0527***
	(.0053)	(.005)	(.0057)	(.0052)
Mean of DV	2.1799	2.1799	1.4288	1.4288
	(.0006)	(.0006)	(.0006)	(.0006)
Year FEs	Yes	Yes	Yes	Yes
Location FEs		Yes		Yes
Demographic controls		Yes		Yes
Observations	900209	900209	900209	900209
Panel B	**Top performance rating**		**Top potential rating**	
	(1)	(2)	(3)	(4)
Female	.0181***	.0061	-.0143***	-.0137***
	(.0044)	(.0043)	(.0017)	(.0019)
Mean of DV	.2496	.2496	.0446	.0446
	(.0005)	(.0005)	(.0002)	(.0002)
Year FEs	Yes	Yes	Yes	Yes
Location FEs		Yes		Yes
Demographic controls		Yes		Yes
Observations	900209	900209	900209	900209

This table reports regressions of Nine Box performance and potential ratings on the female indicator and other control variables for worker demographics, fiscal year fixed effects, and location fixed effects. 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 to 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)	(0.251)	(0.276)
Potential rating				
2=Med	10.61***	10.52***	7.343***	6.838***
	(0.294)	(0.292)	(0.303)	(0.322)
3=High	20.70***	19.50***	14.66***	13.57***
	(0.865)	(0.864)	(0.864)	(0.881)
Performance rating				
2=Med		6.856***	6.358***	5.921***
		(0.329)	(0.332)	(0.379)
3=High		10.09***	10.71***	10.38***
		(0.417)	(0.421)	(0.480)
Log tenure			-1.919***	-1.582***
			(0.142)	(0.163)
Demographic FEs	Yes	Yes	Yes	Yes
Fiscal year FEs	Yes	Yes	Yes	Yes
Location FEs				Yes
Observations	900209	900209	900209	900209

This table replicates Table 2, with the addition of controls for potential rating indicators (the omitted category is 1=Low). By comparing the coefficient on the female indicator in this table with the corresponding indicator in in Table 2, we estimate the fraction of the gender gap in promotions that can be explained gender differences in by 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	(1)	(2)	(3)	(4)
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)
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	(5)	(6)	(7)	(8)
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)
Year FEs	Yes	Yes	Yes	Yes
Demographic controls		Yes		Yes
Location FEs		Yes		
Observations	586338	586338	5222	5222

Panel A reports regressions of Nine Box performance ratings in the following fiscal year on the female indicator, indicators for the worker's potential and performance ratings in the current fiscal year, and other control variables for worker demographics, fiscal year fixed effects, and location fixed effects. Columns 1 and 2 use the full sample while Columns 3 and 4 are limited to observations at the worker-month level corresponding to promotion events. Regressions in Columns 3 and 4 do not control for location fixed effects because of the smaller sample size. Panel B is identical to Panel A except that the dependent variable is the Nine Box potential rating in the following fiscal year. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Table 7: Firm Attachment and the Gender Potential Gap

	(1) Potential rating	(2) Potential rating	(3) Potential rating	(4) Potential rating	(5) Potential rating
Female	-0.0854***	-0.0838***	-0.0815***	-0.0797***	-0.0471***
	(0.00559)	(0.00559)	(0.00559)	(0.00560)	(0.00505)
Past leaves		-0.0235***	-0.0191***		
		(0.00396)	(0.00387)		
Future leaves			-0.0267***		
			(0.00297)		
Fiscal year FEs	Yes	Yes	Yes	Yes	Yes
Performance FEs	Yes	Yes	Yes	Yes	Yes
Leave FEs				Yes	Yes
Location FEs					Yes
Demographic controls					Yes
Observations	900209	900209	900209	900209	900209

Table 8: Increasing gender gaps by pay decile

	(1) Performance rating	(2) Potential rating	(3) Log salary	(4) Promoted	(5) Promoted
Female	0.0864***	0.0261	-0.0399***	8.713***	7.968***
	(0.0192)	(0.0181)	(0.00622)	(1.690)	(1.676)
Pay decile	0.0311***	0.0245***	0.103***	-0.782***	-1.169***
	(0.00203)	(0.00213)	(0.000706)	(0.155)	(0.160)
Female × Pay decile	-0.00486	-0.0120***	-0.000464	-2.002***	-1.882***
	(0.00337)	(0.00340)	(0.00115)	(0.298)	(0.295)
Fiscal year FEs	Yes	Yes	Yes	Yes	Yes
Location FEs	Yes	Yes	Yes	Yes	Yes
Demographic controls	Yes	Yes	Yes	Yes	Yes
Worker ratings FEs					Yes
Observations	160232	160232	160029	160232	160232

This table examines how the gender gaps in ratings, compensation, and promotions vary with pay decile. Pay decile is measured using the ranking of salary within each fiscal year. Column 5 include controls for worker ratings fixed effects, including performance rating indicators and potential rating indicators. All other variables are as defined in previous tables. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Table 9: Variation by manager gender

	(1) Performance rating	(2) Potential rating
Female	0.0157***	-0.0582***
	(0.00564)	(0.00582)
Manager female	-0.0200***	-0.0286***
	(0.00695)	(0.00732)
Female × Manager female	-0.000778	0.0193**
	(0.00945)	(0.00970)
Fiscal year FEs	Yes	Yes
Location FEs	Yes	Yes
Demographic controls	Yes	Yes
Observations	885353	885353

This table examines how the gender gaps in ratings, compensation, and promotions vary with the gender of the worker's immediate manager. Manager female is an indicator for whether the manager is female. All other variables are as defined in previous tables. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Table 10: Variation by manager ratings

	(1) Performance rating	(2) Potential rating	(3) Log salary	(4) Promoted	(5) Promoted
Female	0.0449**	0.00281	-0.135***	1.110	0.861
	(0.0187)	(0.0189)	(0.0123)	(1.264)	(1.254)
Manager potential (1-3)	0.0237***	0.0374***	0.0584***	0.689**	0.333
	(0.00374)	(0.00400)	(0.00261)	(0.272)	(0.271)
Manager performance (1-3)	0.0543***	0.0364***	0.0303***	1.075***	0.571*
	(0.00449)	(0.00465)	(0.00297)	(0.316)	(0.313)
Female × Manager potential (1-3)	-0.0167***	-0.0122**	-0.00346	-0.0327	0.123
	(0.00566)	(0.00601)	(0.00387)	(0.402)	(0.399)
Female × Manager performance (1-3)	-0.00131	-0.0156**	-0.00675	-0.911**	-0.794*
	(0.00668)	(0.00670)	(0.00429)	(0.454)	(0.451)
Fiscal year FEs	Yes	Yes	Yes	Yes	Yes
Location FEs	Yes	Yes	Yes	Yes	Yes
Demographic controls	Yes	Yes	Yes	Yes	Yes
Worker ratings FEs					Yes
Observations	829429	829429	828417	829429	829429

This table examines how the gender gaps in ratings, compensation, and promotions vary with the performance and potential ratings of the worker's immediate manager. Manager potential and manager performance are variables equal to 1, 2, or 3, representing the manager's Nine Box potential and performance ratings, respectively. All other variables are as defined in previous tables. Standard errors are clustered by worker. *, **, and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

TABLE 11: PROMOTION RATES AND PROMOTED WORKER PERFORMANCE UNDER COUNTERFACTUAL POLICIES

	Promotion rates		Expected next performance rating among promoted		
	(1) among men	(2) among women	(3) full sample	(4) true unpromoted sample	(5) 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 future performance ratings under the firm's current promotion policy and counterfactual promotion policies. Details are provided in Section 6. Columns 1 and 2 provide the counterfactual expectations of promotion rates for men and women, respectively. Column 3 provides expectations of the next year's performance ratings, weighted by the current year's promotion probabilities, among all workers. Column 4 does the same, but for workers who were not promoted in the true sample. Column 4 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. 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 1's and 2's when estimating predicted promotion rates. Counterfactual 3 adds one to the potential ratings of women who receive 1's and 2's only for women who receive performance ratings of 3. Bootstrapped standard errors clustered on the worker are parentheses.

8 Appendix Tables

Appendix Figure A1: Nine box ratings and labels

		Performance		
		1 (Low)	2 (Medium)	3 (High)
Potential	3 (High)	**Box 6** New	**Box 3** Delivering, strong potential	**Box 1** High performing, top talent
	2 (Medium)	**Box 8** Potential mismatch	**Box 5** Delivering, promotable	**Box 2** High performing, promotable
	1 (Low)	**Box 9** Underperforming,	**Box 7** Delivering	**Box 4** High performing critical resource