RIETI Discussion Paper Series 25-E-098

Supervisor Accuracy in Subjective Evaluations and Employee Careers

KANAYAMA, Hayato
Waseda University

KAWATA, Yuji
Waseda University

KITAGAWA, Ritsu
Columbia University

RIETI
Research Institute of Economy, Trade & Industry, IAA

The Research Institute of Economy, Trade and Industry
https://www.rieti.go.jp/en/

RIETI Discussion Paper Series 25-E-098
October 2025

Supervisor Accuracy in Subjective Evaluations and Employee Careers*

Hayato KANAYAMA (Waseda University)
Yuji KAWATA (Waseda University)
Ritsu KITAGAWA (Columbia University)

Abstract

We investigate how the accuracy of supervisors' performance evaluations affects employee careers. We develop a simple model in which a supervisor receives a noisy signal of an employee's performance and submits a subjective rating based on that signal. The model predicts that more accurate supervisors generate greater dispersion in their rating scores across subordinates, leading to more promotions. Using personnel records from a large manufacturing firm, we identify supervisors with higher rating dispersion as more accurate raters and estimate the effect of being assigned to them. Consistent with the model's prediction, we find that employees assigned to accurate raters are promoted at higher rates. We also show that supervisors who drink alcohol and those hired more recently tend to be more accurate raters.

Keywords: Subjective performance evaluation, Promotion, Supervisor
JEL classification: M5

*This study is conducted as a part of the Project "Human Capital Investment, Role of Management, and Productivity" undertaken at the Research Institute of Economy, Trade and Industry (RIETI). The draft of this paper was presented at the RIETI DP seminar. We would like to thank participants of the RIETI DP Seminar for their helpful comments. We are indebted to Guido Friebel for inspiring conversations that motivated this study. We are grateful to Yasushi Asako, Stéphane Bonhomme, Laura Caron, Haoge Chang, Ashley Craig, Zoë Cullen, Ryosuke Fujitani, Jumpei Hamamura, Takuma Kamada, Hiroyuki Kasahara, Munechika Katayama, Takao Kato, Daiji Kawaguchi, Ray Kluender, Ayako Kondo, Daiki Nagata, Fumio Ohtake, Erika Mina Okada, Hiroko Okudaira, Hiroshi Ono, Shingo Oue, Shawn Park, Andrea Prat, Hitoshi Shigeoka, Feng Tian, Kozo Ueda, David Weinstein, Shintaro Yamaguchi, and participants at AASLE 2023 Conference, JEA 2024 Spring Meeting, CTW, COPE 2025, AEM Workshop, NBER Japan Project Meeting 2025, SIOE 2025, and Hitotsubashi ICS Seminar for their valuable feedback. We also thank Hideo Owan for granting data access and advising us as the project leader at RIETI. We are especially grateful to the anonymous company that provided the data. The views expressed are those of the authors and do not necessarily reflect those of RIETI, the company providing the data, or any other affiliated organizations. All results have been reviewed to ensure that no confidential information is disclosed. Kanayama gratefully acknowledges financial support from JST SPRING (JPMJSP2128). Kitagawa gratefully acknowledges financial support from the Sylff Association (SRG #2456) and conference travel support provided by the Center on Japanese Economy and Business (CJEB) at Columbia Business School.

1 Introduction

Supervisors play various instrumental roles in organizations, and one of them is employee evaluation. They are often responsible for assessing subordinates' performance, which can have lasting effects on their career trajectories. From the firm's perspective, accurate evaluations are essential for allocating talent efficiently and reinforcing incentive systems (Frederiksen et al., 2017). However, because evaluations are often subjective and shaped by supervisors' own perceptions, experiences, and interpersonal interactions, they can vary in accuracy. Understanding what accounts for variation in evaluation accuracy across supervisors is therefore critical for improving organizational outcomes.

Subjective evaluations are widely used in modern workplaces. This is because employee performance is often difficult to objectively measure, as jobs typically involve complex, multidimensional, and team-based tasks that do not lend themselves to simple metrics (Baker et al., 1994). In such settings, supervisors' subjective assessments serve as a key input for personnel decisions, including promotions, bonuses, and role assignments (Frederiksen et al., 2017). While subjective evaluations can incorporate valuable contextual information, they are also prone to substantial heterogeneity across supervisors, much of which remains unexplained and may reflect differences in accuracy, bias, or managerial ability.

In this study, we focus on the accuracy of supervisors' evaluations and examine how variation in this accuracy affects employee promotion outcomes. We begin by developing a simple model in which supervisors receive noisy signals of subordinate performance and predict that more accurate supervisors generate greater dispersion in assessment scores. Guided by this prediction, we identify accurate raters as those with higher rating dispersion. Using personnel records from a large manufacturing company, we show that employees assigned to accurate raters are promoted at higher rates than those reporting to less accurate ones. We also provide supporting evidence for additional model predictions and examine the characteristics associated with accurate raters.

Our model is one in which the supervisor receives a noisy signal of her subordinate's performance or ability. We assume that the supervisor minimizes the expected squared difference between the true performance and her assessment given the signal she receives. The model predicts that supervisors who receive accurate signals exhibit more dispersed assessment scores across subordinates. The intuition is that when signals are more precise, supervisors rely more heavily on them, whereas when signals are less precise, they tend to base their assessments on the prior mean of the subordinate performance distribution. If the firm's promotion policy can be approximated as a threshold rule, the model further predicts that employees reporting to more accurate raters are promoted at higher rates. The

model also predicts that accurate raters assess high-performing employees more favorably and low-performing employees more critically than less accurate raters. As a result, employees promoted by accurate raters are, on average, of higher quality than those promoted by less accurate raters. We also show that these predictions are likely to hold across different specifications of the supervisor payoff function.

To test our predictions, we use personnel records from a large manufacturing company. We restrict our sample to male employees, who account for over 80 percent of the firm's workforce, as female employees typically follow different career trajectories. The data spans 2006 to 2019, and the main estimation sample includes 6,512 employee-year observations. The firm's assessment records provide annual evaluation scores and identify each employee's direct supervisor. We compute the variance of annual subjective assessment scores across subordinates for each supervisor. To isolate variation attributable to supervisors' ability to differentiate among subordinates, we first residualize assessment scores by partialing out employee, division, supervisor, and year fixed effects. This procedure controls for individual ability, organizational context, leniency differences, and time trends. We use the square root of this variance as a proxy for supervisor accuracy in subjective evaluations and find substantial heterogeneity: switching from a supervisor at the 10th percentile to one at the 90th percentile is associated with an approximately 7.2 percentage-point increase in promotion probability.

We estimate an event-study model and find consistent results. In our event-study estimation, we define the event as being assigned a supervisor whose accuracy is above the median and also control for the effect of supervisor switches per se. The event-study estimates indicate that employees are 10.0 percentage points more likely to be promoted three years after being assigned to an accurate rater. By running a quantile regression, we also show that high-performing employees receive more favorable assessments, and low-performing employees more critical ones, when evaluated by accurate raters compared to less accurate raters. We also find that employees promoted under accurate raters are of higher quality than those promoted under less accurate raters, as measured by the number of subsequent promotions they receive after the focal supervisor. These empirical results are consistent with the model predictions.

We also examine the characteristics associated with accurate raters. We find that supervisors identified as accurate raters are more likely to be frequent/heavy drinkers, which may proxy for greater social skills or extraversion. Interestingly, however, we do not find a significant advantage for employees who drink when assigned to a supervisor who also drinks, which contrasts with the findings of Cullen and Perez-Truglia (2023). Supervisors who consistently give higher ratings are less likely to be accurate, plausibly because they

fail to differentiate employee performance when scores are shifted toward the upper limit. Notably, supervisors' own competence, measured by their promotion speed as used in Minni (2023), does not explain variation in evaluation accuracy.

This paper contributes to several key strands of literature in personnel and organizational economics. First of all, it builds on the classical work that examines subjective evaluations as a response to incomplete performance measurement in modern, multidimensional jobs (Baker et al., 1994; Holmstrom and Milgrom, 1991; Levin, 2003). When some tasks are essential but difficult to measure, employees may underprovide effort unless incentives are designed to account for them. Subjective evaluations offer one solution by incorporating harder-to-measure tasks into incentive schemes, a point supported by both theoretical and empirical studies (e.g., Baker et al., 1994; Bushman et al., 1996; Hayes and Schaefer, 2000; Gibbs et al., 2004; Fuchs, 2007; Frederiksen et al., 2017; Takahashi et al., 2021). At the same time, the use of subjective evaluations can introduce distortions that undermine organizational effectiveness. Because such assessments rely on individual judgment rather than objective metrics, they are vulnerable to biases such as favoritism, centrality bias, leniency, or discrimination (e.g., Medoff and Abraham, 1980; Prendergast and Topel, 1993, 1996; Elvira and Town, 2001; Levin, 2003; MacLeod, 2003; Thiele, 2013; Frederiksen et al., 2020; Benson et al., 2024). These distortions can weaken the link between performance and rewards, potentially reducing employee motivation and misallocating talent. Our study contributes to this strand of literature by documenting how evaluation accuracy translates to employee career outcomes.

Second, we also contribute to the emerging literature that sheds light on supervisor heterogeneity and behavior in employee evaluation. Although performance evaluation plays a central role in incentive contracts, relatively little attention has been paid to the delegation of employee ratings to supervisors and the implications of heterogeneity across supervisors. An important recent contribution is Frederiksen et al. (2020), who document substantial heterogeneity across supervisors in subjective employee evaluations. While their focus is on differences in supervisors' average rating levels, which they refer to as "leniency bias," our analysis centers on heterogeneity in rating accuracy, measured by the dispersion in rating scores across supervisors. Closest to our approach is Kampkötter and Sliwka (2018), who show that greater dispersion in evaluations is associated with higher subsequent bonus payments. Other notable contributions to this literature include Kawaguchi et al. (2016), who document supervisor bias in subjective performance evaluations at a large manufacturing firm, Takahashi et al. (2021), who show by using data from a car-sales company that workers' reactions to unexpectedly low evaluations vary with supervisor experience, and Haegele (2024), who finds that managers incentivized to hoard talent rate subordinates lower than

deserved. Our findings complement the literature by highlighting the substantial impact of supervisor rating accuracy on employee career outcomes.

Third, our findings are also relevant to the more broadly defined literature on delegated assessments by managers. For example, Hoffman et al. (2018) show that hiring managers who override algorithmic recommendations often select candidates who subsequently leave their jobs early, suggesting inefficiencies in managerial discretion. Similarly, Shukla (2025) documents that hiring managers systematically discriminate against candidates from lower social classes in the informal screening stage, which is supposed to assess softer information such as candidates' fit. In contrast, Wu and Liu (2020) find that delegating hiring authority to local managers improves the average productivity of new hires and enhances store-level performance. Relatedly, Friebel et al. (2024) show that regional managers possess valuable private information that helps both researchers and the firm predict which stores will benefit most from a new HR policy. Taken together, these studies underscore both the promise and the risk of relying on subjective managerial assessments. Our study extends this literature by quantifying supervisor accuracy in internal performance evaluations and showing its impact on employee career outcomes, thereby highlighting the consequences of delegated discretion in employee performance evaluation.

Fourth, we contribute to the growing literature on the role of middle managers and first-line supervisors. Lazear et al. (2015) show that supervisor quality, measured by value-added to subordinate productivity, can vary substantially even within the same firm. Subsequent studies complement their findings by documenting the importance of specific managerial skills, abilities, and behaviors, such as technical competence, communication, evaluation, interpersonal skills, energy efficiency control, talent allocation, training, and social interactions, in shaping managerial quality (e.g., Artz et al., 2017; Kuroda and Yamamoto, 2018; Frederiksen et al., 2020; Hoffman and Tadelis, 2021; Metcalfe et al., 2023; Cullen and Perez-Truglia, 2023; Minni, 2023; Asuyama and Owan, 2024; Diaz et al., 2025; Macdonald et al., 2025). Our work aligns with this direction by emphasizing evaluation accuracy as an important and measurable dimension of managerial quality that affects subordinate outcomes.

The rest of the paper is structured as follows: Section 2 explains our conceptual framework. Section 3 describes the data and organizational context of the company. Section 4 presents our empirical strategy and results. Section 5 concludes.

2 Conceptual Framework

To formalize our idea about supervisors' evaluation decisions, we present a simple conceptual framework, inspired by Kawaguchi et al. (2016) and Frederiksen et al. (2020). We consider

a model in which a supervisor evaluates the performance of her subordinate.[1] Assume that subordinate performance is normally distributed, $q \sim \mathcal{N}(\overline{q}, \sigma_q^2)$. The supervisor receives a noisy signal about the subordinate's ability, $\widehat{q} = q + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma_s^2)$. Assume q and ε are independent. Based on this signal $\widehat{q}$, the supervisor submits a rating score r. Assume that submitting an assessment deviating from the true ability is costly for the supervisor. Her utility is given by the expected negative value of the squared error of her evaluation, conditional on the signal:

$$\mathsf{E}\left[-(r-q)^2 \mid \widehat{q}\right].$$

The intuition behind this assumption is that overestimating a subordinate may harm the supervisor's reputation if the promoted employee performs poorly in a role he did not merit. On the other hand, underestimating a subordinate can also damage the supervisor's reputation, as the employee may complain to the HR department about unfair treatment. In Appendix B, we explore alternative specifications of the supervisor payoff, such as payoffs that exhibit leniency, aversion to negative feedback, and talent hoarding, and show that our main theoretical message remains largely unchanged.

The supervisor's problem is

$$\max_{r} \quad \mathsf{E}\left[-(r-q)^2 \mid \widehat{q}\right]. \tag{1}$$

By the standard signal extraction exercise, we can show that the supervisor's optimal assessment r^* is given by

$$r^* = \frac{\sigma_s^2}{\sigma_q^2 + \sigma_s^2}\overline{q} + \frac{\sigma_q^2}{\sigma_q^2 + \sigma_s^2}\widehat{q}. \tag{2}$$

We introduce two types of supervisors, indexed by $s \in \{A, N\}$. Here, A stands for accurate-rating supervisors and N stands for non-accurate-rating supervisors. Accordingly, assume $\sigma_A^2 < \sigma_N^2$. The intuition of our specification is that accurate-rating supervisors are more effective in exerting effort to gather accurate information about employees. Assume that the firm's promotion policy can be approximated by a threshold rule such that the subordinate is promoted when $r \geq t$ for some given $t \geq \overline{q}$.[2] The assumption $t \geq \overline{q}$ is empirically

[1] In this section, we keep the discussion as simple as possible and refer the reader to Appendix A for a more explicit treatment of the employer's problem.

[2] Since subjective evaluation is a comprehensive measure of performance, it should capture most of the relevant information used to predict promotion outcomes. If other factors also influence promotion, we can incorporate them into the decision rule as a residual component, and write the promotion threshold rule as $r + \eta \geq t$, where η represents the unobserved residual. As long as r and η are independent, this threshold approximation remains valid. Another concern would be the case where the firm is aware of heterogeneity in supervisors' evaluation accuracy, reflected in differences in σ_s^2. Then, the firm may optimally adjust the

plausible because, in most organizations, fewer than half of employees are promoted to the next level in any given cycle. Note that $\widehat{q} \sim \mathcal{N}(\overline{q}, \sigma_q^2 + \sigma_s^2)$. Since $r^* \sim \mathcal{N}\left(\overline{q}, \frac{\sigma_q^4}{\sigma_q^2+\sigma_s^2}\right)$, the probability that the subordinate is promoted is

$$\mathsf{Pr}(r^* \geq t) = 1 - \Phi\left(\frac{\sqrt{\sigma_q^2 + \sigma_s^2}}{\sigma_q^2}(t - \overline{q})\right), \tag{3}$$

where Φ is the normal cumulative distribution function.

Let r_A^* and r_N^* denote r^* under $s = A$ and $s = N$, respectively. We can derive the following predictions. All proofs are deferred to Appendix D.

Prediction 1 (Dispersed assessments) *Accurate raters receive more precise signals and therefore rely more heavily on those signals, producing more dispersed evaluations. Less accurate raters rely more on priors, leading to compressed scores around the mean. Namely,* $\mathsf{Var}[\, r_A^* \,] > \mathsf{Var}[\, r_N^* \,]$.

Prediction 2 (Higher promotion rate) *Since accurate raters produce a wider distribution of scores centered around the same mean, more employees surpass the promotion threshold at the upper tail and get promoted. Namely,* $\mathsf{Pr}(r_A^* \geq t) > \mathsf{Pr}(r_N^* \geq t)$.

Prediction 3 (Accuracy and amplified differentials) *Accurate raters adjust their evaluations more sharply in response to differences in true performance. This leads to greater separation in scores between high- and low-performing employees compared to less accurate raters. Namely, for* $q' > q$, $\mathsf{E}[\, r_A^* \mid q' \,] - \mathsf{E}[\, r_N^* \mid q' \,] > \mathsf{E}[\, r_A^* \mid q \,] - \mathsf{E}[\, r_N^* \mid q \,]$.

Prediction 4 (Higher quality among promotees) *Because accurate raters' scores better reflect true ability, the set of employees they promote includes fewer false positives and thus has higher average ability than the set promoted by less accurate raters. Namely,* $\mathsf{E}[\, q \mid r_A^* \geq t \,] > \mathsf{E}[\, q \mid r_N^* \geq t \,]$.

We treat the first prediction as a premise that guides our empirical strategy for identifying accurate raters. In our empirical analysis, supervisors with greater dispersion in assessment scores are identified as accurate raters. The remaining predictions provide testable implications based on this definition of accurate raters. If dispersion captures evaluation accuracy, employees reporting to accurate raters should be promoted more often. This is not because of greater leniency but because precise signals allow stronger performers to stand out and surpass the promotion threshold. At the same time, because all supervisors share the same

promotion threshold to supervisor type. In Appendix C, we show that it is not the case in our setup.

prior and truthfully report their signals, average scores should remain similar across supervisor types. The model also predicts that accurate raters respond more strongly to differences in true performance, generating sharper score differentials between high- and low-ability subordinates. As a result, the set of employees promoted by accurate raters should have higher average ability than those promoted by less accurate raters. These implications form the basis for our empirical analysis.

3 Data and Organizational Context

3.1 General Background

Our dataset combines several primary sources of personnel records from the firm. First, we use annual evaluation records that identify each employee's direct supervisor and include performance assessments provided by that supervisor. Second, we use pay-grade records. Third, we use annual engagement survey responses, which capture employees' self-reported attitudes toward their work and workplace environment. Finally, we also utilize annual health check records, which include biometric indicators and lifestyle information. Because employers in Japan are legally required to offer health examinations, coverage is nearly universal. These records also include questions on drinking, smoking, and exercise habits, which we use as potential predictors of supervisor accuracy.

We restrict our sample to non-managerial male permanent employees in the lowest pay-grade range, which comprises five grades. We define a promotion as any upward movement in an employee's pay grade. The average annual promotion rate in this group is 13 percent, meaning that approximately 13 out of every 100 employees are promoted each year. Our focus is on frontline supervisors who are in charge of directly supervising the non-managerial workforce.[3] At the study company, frontline supervisors are the ones who submit annual performance evaluations of their direct reports. Employees take their annual evaluations seriously because they are a crucial determinant of their bonuses and promotions. We can identify the supervisor for each employee in December, when annual evaluations are submitted to the company's human resource management system.

The firm periodically rotates employees across different divisions, which enables us to observe a sufficient number of supervisor transitions driven by both supervisor and employee transfers. Personnel decisions are centralized to a great extent, and employees do not have a voice in where and with whom to work. An average employee in our sample reported to 3.87 different supervisors over the observation periods. The average span of control ranges

[3]Most supervisors in this company have managerial titles.

from 2.8 to 4.4 employees across years, and the average supervisor-subordinate tenure is 26.3 months.

Our sample includes both white-collar and blue-collar workers. More specifically, it includes those with administrative, sales, R&D, and production workers. Employees with administrative roles are those who work in such as accounting, human resource, and public relations divisions. Those with sales and R&D roles are those who are assigned to sales and R&D divisions, respectively. Those with production roles are those who work in the factories and directly engage in the manufacturing process. We observe this job-functional information for each employee-year observation up to the fiscal year 2018. Unfortunately, this information is not available for the fiscal year 2019. From 2006 to 2018, administrative, sales, R&D, and production roles account for 17.9 percent, 22.2 percent, 28.0 percent, and 31.9 percent of our employee-year observations, respectively. Approximately 70 percent of the unique employees in our sample never experienced a change in their job functions, and, in estimation, most variations will be absorbed by the two-way fixed effects. The attrition rate in this company is quite low, ranging from 1.5 percent to 5 percent depending on the year. The average age and tenure are 40.9 years old and 17.1 years, respectively.

There are five pay grades for non-managerial employees, labeled from the lowest to highest as Rank 1 through Rank 5. Employees who advance beyond Rank 5 transition into managerial roles and become exempt from hourly wage regulations. Pay grades determine both base salary and the applicable range for performance-based bonuses. Each pay grade is associated with a schedule of bonus points, which determine bonus amounts in combination with evaluation results. Figure 1 shows the distribution of bonus points by evaluation grade, with colors indicating employees' pay grades. Bonus points increase monotonically with evaluation grade. Within each evaluation grade, employees at higher pay grades tend to receive more bonus points, reflecting the firm's policy of linking rewards to both performance and position.

Figure 2 shows how employee pay grades evolve with tenure. At the beginning of their careers, nearly all employees start at grade Rank 1. Within the first five years, most transition quickly to Rank 2, reflecting the firm's almost automatic promotion practices. After that point, the pace of promotion slows and becomes more varied across individuals. As tenure increases, employees gradually move into higher grades such as Rank 3, Rank 4, Rank 5, and eventually into positions above Rank 5. By around 15 years of tenure, a large share have reached Rank 5 or higher, but some remain in lower grades even after longer careers. This pattern suggests that early promotions are common and largely tenure-based, while later promotions appear more dependent on performance or other criteria.

For our main empirical exercise, we define promotions as any increase in one's pay

grade from the preceding year. This definition captures both routine grade progressions and performance-based advancements. Since pay grade is a key determinant of base salary and bonus eligibility, upward movement reflects meaningful career progression within the firm. We focus on transitions within the non-managerial grade range (Rank 1 to Rank 5), where most employees spend the bulk of their early to mid-careers. Demotions are negligibly rare in our sample, with only a few observed cases over the entire period. As a result, we focus exclusively on upward grade transitions when analyzing promotion patterns. We exclude employees who appear in the data for the first time, as we cannot define their promotion outcome without information from the preceding year.

3.2 Performance Evaluation

The performance evaluation system remained broadly consistent throughout the observation period, employing a seven-point rating scale in all years. However, the labeling of the scale changed in 2011. Before 2011, ratings ranged from highest to lowest as follows: S, A1, A2, B1, B2, C, and D. From 2011 onward, the labels were revised to SS, S, AA, A, B, C, and D, again in descending order. In both periods, the fourth level from the top (B1 before 2011 and A after 2011) serves as the baseline specified by the firm as the standard for satisfactory performance. Figure 3 compares the distribution of ratings before and after the labeling change. The two distributions are very similar, suggesting the label change did not affect rating behaviors. For the empirical analysis, we use a standardized score for each employee-year rating. Specifically, each score is normalized by subtracting the fiscal-year mean and dividing by the fiscal-year standard deviation. Figure 4 shows the distribution of these standardized scores.

The performance ratings used in our analysis are those employed by the firm in determining employee bonuses, making them a consequential and salient component of the personnel system. These evaluations are submitted annually by each employee's direct supervisor, typically a frontline manager within the same work unit who observes the employee's day-to-day activities. Because the evaluations are directly tied to financial rewards and are used to inform promotion decisions, supervisors are expected to take them seriously and provide assessments that reflect actual job performance. Given their institutional role and practical consequences, these scores are well suited for our empirical analysis. They represent the key channel through which supervisors influence career outcomes, and they are available consistently across employees, supervisors, and years. Moreover, the structured and repeated nature of the evaluation process allows us to construct supervisor-specific measures of evaluation behavior such as rating dispersion while controlling for confounding factors

like employee ability and organizational context. This makes the evaluation score a credible object of study for examining how variation in supervisor accuracy affects employee careers.

4 Empirical Strategy and Results

4.1 Measuring Supervisor Accuracy

Our measurement of supervisors' accuracy in subjective evaluations is guided by the conceptual framework presented in Section 2. According to Prediction 1, supervisors who receive more precise signals about employee performance, that is, more accurate raters, place greater weight on those signals when forming evaluations. As a result, their assessment scores exhibit greater dispersion. In contrast, less accurate raters tend to rely more heavily on prior beliefs, producing compressed scores clustered around the mean. We use this predicted variation in rating dispersion as the basis for identifying evaluation accuracy.

Following the model's prediction, we use the variance of each supervisor's empirical distribution of assessment scores as a proxy for their evaluation accuracy. Supervisors whose scores exhibit greater dispersion are interpreted as relying more on informative signals and are therefore classified as more accurate raters. Let i index employees and t denote the year of observation. Define d to be the mapping from employee i to his or her division $d(i,t)$ in year t and $s(i,t)$ to be the mapping from employee i to his supervisor s in t, respectively.

Now, we estimate the following regression equation to isolate supervisor-specific variation in assessment scores:

$$y_{i,t} = \alpha_i + \delta_{d(i,t)} + \gamma_{s(i,t)} + \tau_t + \eta_{i,t}, \tag{4}$$

where $y_{i,t}$ denotes the subjective performance score received by employee i in year t. The term α_i is employee fixed effects capturing time-invariant differences in employee competence. We also control for division fixed effects, represented by $\delta_{d(i,t)}$, that absorb differences across organizational units. In addition, we include $\gamma_{s(i,t)}$, which represents supervisor fixed effects, which capture systematic differences in the levels such as leniency, and τ_t indicates year fixed effects accounting for time trends in scoring practices or firm-level policies.

To exclude variation driven by subordinate competence, supervisor fixed effects, and divisions, we residualize assessment scores using the estimates from equation (4).

$$\widehat{\eta}_{i,t} = y_{i,t} - \widehat{\alpha}_i - \widehat{\delta}_{d(i,t)} - \widehat{\gamma}_{s(i,t)} - \widehat{\tau}_t. \tag{5}$$

Let $S_{s,t}$ denote the set of employees supervised by supervisor s in year t, that is, $S_{s,t} = \{i : s(i,t) = s\}$. We use the following quantity as a proxy for supervisor s's accuracy in employee

performance assessment:

$$\widehat{\mathsf{Var}\left[r_s^*\right]} = \frac{1}{\sum_{t=1}^{T}|S_{s,t}|}\sum_{t=1}^{T}\sum_{i\in S_{s,t}}\left(\widehat{\eta}_{i,t}\right)^2, \tag{6}$$

where $\widehat{\eta}_{i,t}$ is the residual from Equation (4). This expression captures the within-supervisor variance in residualized assessment scores, consistent with the prediction from our model that more accurate supervisors produce more dispersed evaluations.[4]

Figure 5 displays the distribution of evaluation variance across supervisors using two different measures. The red bars represent the naive variance, calculated directly from raw assessment scores. The blue bars represent the estimated variance based on residualized scores, which remove variation associated with employee, supervisor, division, and year fixed effects. Compared to the naive measure, the estimated variance is more tightly concentrated, suggesting that some of the variation in raw scores may reflect systematic factors unrelated to evaluation accuracy, such as differences in leniency or subordinate composition. The estimated measure provides a more refined indicator of the degree to which supervisors differentiate among their subordinates. The summary statistics of estimated supervisor accuracy in subjective evaluations is reported in Table 1.

Figure 6 plots the distribution of evaluation scores, standardized within year, separately for employees reporting to accurate and non-accurate raters. Accurate raters are defined as those whose estimated evaluation variance is above the median. Standardization removes year-specific shifts in evaluation levels and spreads due the system reform, allowing comparison across years. While both distributions are roughly centered around zero, evaluations from accurate supervisors exhibit greater dispersion. This pattern is consistent with the model's prediction that higher variance in scores reflects greater accuracy in performance evaluation.

To summarize, we use within-supervisor variance in residualized evaluation scores as a proxy for evaluation accuracy, based on the model's prediction that more accurate raters produce more dispersed assessments. This measure accounts for heterogeneity in employee ability, supervisor leniency, and structural differences across divisions and years. Our empirical analyses that follow will use this proxy to examine whether supervisors who provide more differentiated evaluations also promote higher-quality employees and whether evaluation accuracy is systematically associated with observable supervisor characteristics.

[4]We also calculate the variance using the leave-one-out approach. The main result with this alternative proxy is reported in Appendix G.

4.2 Supervisor Accuracy and Promotions

4.2.1 Event-Study Estimation

We now test Prediction 2, which states that employees are promoted at higher rates under accurate raters, by regressing promotion outcomes on supervisor evaluation accuracy. We begin by estimating the following event-study model that exploits quasi-random variation in supervisor-employee assignments. Let $T_{i,t}$ be the indicator function that takes 1 if employee i experiences a supervisor transition in year t. We define $D_{i,t}$ to be the treatment variable that takes 1 if employee i starts reporting to an accurate-rating supervisor in year t defined. Accurate-rating supervisors are defined as those whose estimated accuracy is above the median.

When an employee experiences more than one supervisor transition over the observation periods, we focus on the first one. We also restrict the sample to employees observed within three years of their first supervisor switch and who continued working under the same supervisor afterward. This ensures that we capture the immediate effects of supervisor changes while avoiding confounding influences from repeated or staggered switches. Limiting to employees who remained with the new supervisor allows us to isolate the effect of supervisor accuracy without additional variation in reporting relationships. This restriction leaves us with 8,141 employee-year observations. Since the first observations are omitted because we cannot see whether employees were promoted from the previous year, the promotion analyses use 6,512 observations. We use this estimation sample for the empirical exercises that follow.[5]

The event-study model we estimate is:

$$y_{i,t} = \sum_{k \in \mathcal{K}} \beta_k^D D_{i,t-k} + \sum_{k \in \mathcal{K}} \beta_k^T T_{i,t-k} + \alpha_i + \tau_t + \varepsilon_{i,t}, \tag{7}$$

where the fixed effects of employees (α_i), and years (τ_t) are controlled. On the left-hand side, $y_{i,t}$ is the dummy variable indicating whether employee i is promoted in year t. We also control for the squares of age and tenure. We set the window within three years from the event, i.e., $\mathcal{K} = \{-3, \ldots, -2, 0, 1, 2, \ldots, 3\}$. The parameter β_k^D captures the effect of being assigned to an accurate-rating supervisor among employees who experience a supervisor transition. Our identification relies on the parallel trends assumption, which we assess indirectly using event-study graphs. Substantively, the plausibility of this assumption is supported by the centralized nature of personnel transfers. Neither employees nor supervisors have much discretion in determining their assignment, reducing concerns about endogenous matching.

[5] We present and discuss the estimates obtained from the full sample in Appendix F.

We can also indirectly corroborate our assumption by showing the parallel trends before the event. We cluster our standard errors at the employee and supervisor levels.

Figure 7 presents event-study estimates of the effect of being assigned to an accurate-rating supervisor on the probability of promotion. The horizontal axis shows years relative to the supervisor switch. The estimates before the switch (years –3 to –1) are close to zero and statistically insignificant, supporting the parallel trends assumption. In the three years following the switch, promotion probability rises by approximately 10.0 percentage points and the estimate is statistically significant, providing evidence of a positive causal effect of accurate-rating supervisors on employee promotions. One possible interpretation of the estimates being significant only in period 3 is that even supervisors with high rating accuracy need time to form confident assessments of their subordinates. Alternatively, it may also simply reflect a lag between accurate performance evaluations and their incorporation into formal organizational decisions, such as promotions.

The results are similar when we employ the leave-one-out specification of the supervisor accuracy measure discussed in Section 4.1. Although cohort-specific effects are less of a concern in our setting, we apply the estimation procedure of Sun and Abraham (2021) to address the potential staggered nature of supervisor transitions. We also estimate the effects by distinguishing transitions from a non-accurate rater to an accurate raters and transitions from accurate raters to non-accurate raters, with an econometric model similar to Cullen and Perez-Truglia (2023) and Minni (2023). For the last two robustness check exercises, though they become less precise and lose statistical significance, the point estimates remain similar. The above three robustness checks are discussed in detail in Appendix G.

4.2.2 Two-Way Fixed-Effects Estimation

The event-study estimation shows that promotion rates rise only after employees start reporting to accurate raters. Next, we estimate a simpler regression equation with two-way fixed effects to facilitate interpretation and exploit the full variation in supervisor accuracy rather than dichotomize it by treating accuracy as a continuous measure. Specifically, we estimate the following simple regression equation:

$$y_{i,t} = \rho\, \widehat{\sigma}_{s(i,t)} + \alpha_i + \tau_t + \varepsilon_{i,t}, \tag{8}$$

where $y_{i,t}, \alpha_i$, and τ_t denote promotion, employee fixed effects, and time fixed effects, respectively. Here, we use supervisor accuracy as a continuous variable and $\widehat{\sigma}_{s(i,t)}$ is defined such that $\widehat{\sigma}_{s(i,t)} = \left(\widehat{\mathsf{Var}[r^*_{s(i,t)}]}\right)^{1/2}$. We cluster our standard errors at the employee and supervisor levels.

Table 2 presents regression results examining the relationship between supervisor evaluation accuracy and employee promotions. The estimates are positive and statistically significant, suggesting that employees evaluated by more accurate raters are more likely to be promoted. From Table 1, the difference in supervisor accuracy between the 10th and 90th percentiles is 0.90. Multiplying this by the coefficient from Column (4) of Table 2, we find that this variation is associated with a 7.2 percentage-point increase in promotion probability ($0.90 \times 0.0808 = 0.072$). Taken together with the event-study regression result in Section 4.2.1, our results are consistent with Prediction 2, which states that employees assigned to accurate-rating supervisors are more likely to be promoted.

4.3 Heterogeneity Across Evaluation Quantiles

Next, we test Prediction 3, which states that accurate-rating supervisors better differentiate between high- and low-performing subordinates by assigning systematically higher ratings to more capable employees and lower ratings to less capable ones than less accurate raters would. To examine this, we investigate whether the effect of supervisor accuracy on employee evaluations is heterogeneous across the distribution of subordinate performance. Specifically, we estimate the following equation:

$$y_{i,t} = \rho_q \, \widehat{\sigma}_{s(i,t)} + \boldsymbol{X}'_{i,t}\boldsymbol{\beta} + \xi_{i,t}, \tag{9}$$

with the restriction that

$$\mathsf{Q}_q[\, \xi_{i,t} \,|\, \widehat{\sigma}_{s(i,t)}, \boldsymbol{X}_{i,t} \,] = 0,$$

where subscript q denotes quantiles, $\mathsf{Q}_q[\, Y \,]$ is defined such that $\mathsf{Q}_q[\, Y \leq \mathsf{Q}_q[\, Y \,]\,] = q$, $y_{i,t}$ indicates the evaluation score of employee i at year t as evaluated by his supervisor, and $\boldsymbol{X}_{i,t}$ is a vector of controls that include employee i's job function, employee i's age, tenure, boss age, and their squares in year t, and a constant. Prediction 3 in our conceptual framework suggests that $\rho_{q'} > \rho_q$ for $q' > q$.

Figure 8 presents the results from a quantile regression of performance evaluations on supervisor accuracy, allowing the relationship to vary across quantiles of employee performance. The horizontal axis represents employee performance quantiles (for example, the 15th, 30th, up to the 90th percentile), and the vertical axis shows the estimated association between supervisor accuracy and evaluation scores at each quantile.

The results are consistent with Prediction 3, as the estimated coefficients are negative in the lower tail and positive in the upper tail of the performance distribution. Compared to less accurate raters, accurate-rating supervisors tend to assign lower evaluations to low-

performing employees and higher evaluations to high performers. This indicates that accurate raters differentiate employee performances more sharply. Notably, the pattern is asymmetric. The positive coefficients in the upper quantiles are larger in magnitude than the negative coefficients in the lower quantiles. One possible explanation is that supervisors may be more reluctant to give very low evaluations, due to concerns about discouraging subordinates or harming team morale.[6] As a result, accurate-rating supervisors may still hesitate to give harsh feedback, even when performance is poor, while being more willing to reward strong performance.

4.4 Quality of Promoted Employees

Finally, we test Prediction 4, which states that employees promoted under accurate-rating supervisors are of higher quality than those promoted under less accurate raters. We examine employee performance after the termination of the focal supervisor-subordinate relationship. Specifically, we compare employees who were promoted under accurate-rating supervisors to those promoted under less accurate raters, focusing on their performance under a new supervisor. If promotions by accurate raters reflect favoritism or non-performance-related factors, we would expect these employees to perform no better or even worse than those promoted by less accurate raters. In contrast, if accurate raters are better informed about employee ability, then employees they promote should outperform those promoted by less accurate raters.

We first restrict the sample to employees who experienced at least one supervisor change during the observation period. Next, we focus on those who were promoted while reporting to the supervisor they received in their first boss switch. We then track the number of subsequent promotions these employees receive after they begin reporting to a new supervisor, that is, following a second boss switch. We use promotion as our outcome measure because it serves as a holistic indicator of employee performance, as used by Minni (2023) for the same reason. This procedure yields a cross section of employees promoted under accurate-rating supervisors and those promoted under less accurate raters.

Using this cross-section sample, we estimate the following quasi-Poisson count model:

$$\ln \mathsf{E}[\,(\#\text{ of promotions})_i \,|\, D_i, \boldsymbol{X}_i\,] = \rho D_i + \boldsymbol{X}_i'\boldsymbol{\beta}, \tag{10}$$

where D_i indicates whether employee i was assigned an accurate rater for the first supervisor

[6] Morita (2025) studies how supervisors balance the developmental benefits of negative feedback against the risk of lowering worker confidence, and shows that those with higher evaluation ability are more likely to suppress negative feedback.

switch and $\boldsymbol{X}_i$ is a vector of controls that include age, tenure, and their squares, when they were promoted. If accurate-rating supervisors effectively differentiate subordinates with high ability from others, then employees promoted under accurate raters should perform better than those promoted under less accurate raters. Accordingly, we predict the coefficient ρ to be positive.

Table 3 presents the estimation result. The key independent variable is an indicator for whether the employee was promoted under a supervisor with high evaluation accuracy. The coefficient on this variable is positive with the average marginal effect being 0.052, suggesting that employees promoted by accurate raters tend to receive slightly more promotions under subsequent supervisors. The direction of the effect is consistent with Prediction 4, which argues that employees promoted under accurate raters are of higher quality.[7]

4.5 Discussion

We have examined how variation in supervisor evaluation accuracy shapes employee careers, focusing on promotions and performance assessments. Building on a simple theoretical model in which a supervisor receives a noisy signal about her subordinate's performance, we have derived several predictions. Our empirical findings are broadly consistent with the model's predictions.

First, we have shown that employees assigned to accurate-rating supervisors have been promoted at higher rates following supervisor transitions, even after accounting for potential confounders through an event-study design that leverages quasi-random assignment. This supports the idea that the accuracy of subjective evaluations plays a meaningful role in career advancement. Notably, this result is aligned with the empirical findings of Kampkötter and Sliwka (2018), who show that more dispersed evaluations are associated with higher bonuses.

Second, we have investigated the relationship between supervisor accuracy and the content of evaluations across the performance distribution. Using quantile regression, we have found that accurate-rating supervisors tend to assign lower scores to low performers and higher scores to top performers, relative to less accurate supervisors. These findings suggest that accuracy is associated with sharper differentiation, rather than uniformly more favorable ratings. The positive associations in the upper tail have been stronger than the negative associations in the lower tail. This pattern is consistent with the idea that supervisors may be more reluctant to issue harsh evaluations, possibly due to concerns about discouraging subordinates or undermining team morale.

[7] This theoretical prediction and its empirical result resonate with Lazear (2004), who shows that the Peter Principle is more pronounced when the measurement error of ability is large.

Third, we have examined the career outcomes of employees who were promoted under accurate-rating supervisors by estimating a quasi-Poisson model of subsequent promotion counts. Among employees who received a promotion while working under a newly assigned supervisor, those promoted by accurate raters experienced higher rates of promotion in the following years, relative to those promoted by less accurate supervisors. Although the result is not statistically significant, it is suggestive that promotions made under accurate-rating supervisors are more predictive of continued career advancement.

A natural question to ask next is who these accurate raters are and what characteristics predict their evaluation accuracy. Understanding the sources of this supervisor heterogeneity can help organizations identify which supervisors are more or less capable of making accurate performance assessments and use this information to guide management practices and decision-making. It is also important to disentangle this heterogeneity into observable characteristics, as our understanding of what drives the value of non-top managers and supervisors remains limited. By doing so, we contribute to the growing literature on the role and effectiveness of middle management. (e.g., Lazear et al. (2015); Hoffman and Tadelis (2021); Friebel et al. (2022); Metcalfe et al. (2023); Friebel et al. (2024)).[8] In the next section, we explore which observable supervisor characteristics are systematically associated with higher evaluation accuracy.

4.5.1 Characteristics of Accurate Raters

Having documented substantial heterogeneity in evaluation accuracy across supervisors, we now turn to examining which observable characteristics are associated with this variation. Identifying predictors of evaluation accuracy can provide insights into the types of supervisors who are more effective at assessing performance and can help organizations make more informed managerial appointments. To this end, we regress our measure of supervisor accuracy on several supervisor characteristics. While the choice of variables is partly constrained by data availability, it is also based on our educated guess as to plausible factors that may influence evaluation accuracy. Those independent variables require explanation in terms of their definitions and the reasons we include them.

First, entry year refers to the year the supervisor joined the company, so higher values indicate more recent hires. It is possible that supervisors find it easier to evaluate subordinates with similar tenure, as they may share comparable experiences or viewpoints. On the other hand, having greater experience may help supervisors better distinguish differences in performance, especially when evaluating less experienced subordinates. Since our sample consists of lower-ranked employees, who tend to be relatively early in their careers, the supervisor's

[8] See Roberts and Shaw (2022) for a comprehensive review.

entry year may influence evaluation accuracy in both ways. A similar idea appears in the findings of Kawata and Owan (2022), who show that workers may be influenced differently by their senior colleagues depending on their own age. This suggests the possibility that proximity in tenure or age could play a role in how supervisors relate to and evaluate their subordinates. For ease of presentation, we refer to supervisors who joined the firm more recently (above the median year of entry) as *younger*.

Second, we define *high-flyers* as supervisors who were promoted to a managerial position or to grade Rank4–Rank5 by the age of 35, which conceivably indicates that the firm regarded them as high-potential employees.[9] This definition follows Minni (2023), who finds that high-flyers tend to be more effective managers. Specifically, she shows that employees reporting to high-flyers advance more quickly in their careers because these supervisors help guide subordinates to roles with better match quality. Her findings suggest that high-flyers may possess stronger ability to observe and assess their subordinates, which motivates the inclusion of this variable. A dummy variable indicating whether the supervisor is a high-flyer is included in the regression.

Third, we include three health-related variables, which come from the mandatory employee annual health check results.[10] First, We include smoking habits as a variable, motivated by Cullen and Perez-Truglia (2023), who find that shared smoking breaks with managers improve subordinates' career outcomes.[11] Their study suggests that smoking may serve as a channel for informal interaction, which can in turn influence the accuracy of supervisors' employee assessments. Second, we include drinking habits for a similar reason, drawing on Wang et al. (2023), who study drinking as a means of business socializing and examine a policy that bans government officials from attending business banquets.[12] Lastly, we also include exercise habits, again, motivated by a similar logic. Like smoking and drinking, exercise may also serve as a socializing tool, as several studies have shown that golfing can provide networking opportunities and contribute to business success (Agarwal et al., 2016; Biggerstaff et al., 2024; Izumi et al., 2024).[13] In addition, these three variables may also

[9]The average tenure for 35-year-old employees is 10.6 years. As shown in Figure 2, approximately 25 percent of employees aged 35 and over meet our definition of high-flyers. This share is comparable to the 29 percent reported by Minni (2023), who define high-flyers as workers promoted by the age of 30.

[10]Health check results are available from 2015 to 2019. For each supervisor, we use the earliest record available during this period.

[11]Smoking status is recorded as yes or no in the health check questionnaire.

[12]Drinking status is assessed in two ways in the health check questionnaire: one item asks about the frequency of alcohol consumption, and another asks about the amount consumed on drinking days. We construct a binary indicator for drinking status based on these responses. The definition of drinkers is provided in Appendix H.

[13]The health check questionnaire asks whether the employee "has engaged in light exercise that works up a slight sweat for 30 minutes or more per session, at least twice a week, for over a year". Responses are coded as yes or no.

capture underlying personality traits or forms of human capital, such as extraversion or social skills, which have been shown to be important attributes for managers (Hansen et al., 2021). These underlying traits may in turn influence how supervisors interact with others and form accurate performance evaluations.

Fourth, we include the supervisor fixed effects estimated from equation (4) that show how high or low the supervisor tends to rate employees overall. As documented by Frederiksen et al. (2017), subjective evaluations are often prone to centrality bias and leniency bias. While our conceptual framework well accounts for centrality bias, which less accurate raters should exhibit, we abstract away from leniency bias. Nonetheless, it is conceivable that supervisors' leniency bias is empirically related to their evaluation accuracy. If supervisors tend to be uniformly lenient, they may have weaker incentives or preference for evaluating subordinates accurately in the first place. Also, more mechanically, when supervisors uniformly give higher ratings, the scores may cluster at the upper end of the scale, leading to lower dispersion due to truncation in the discrete rating system.

Last of all, we use the average supervisor satisfaction score based on the firm's annual employee engagement survey, which reflects how satisfied subordinates are with their supervisor. The survey includes three questions covering different aspects of supervisory behavior: (1) "Do you receive appropriate instructions from your supervisor based on sound judgment?" (2) "Does your supervisor provide you with growth opportunities that match your abilities and personality?" and (3) "Were you satisfied with the beginning-of-year discussion with your supervisor last year?". Responses are recorded on a three- to six-point Likert scales. We standardize these responses by fiscal year and then compute the average score across all subordinates for each supervisor.[14] As prior studies have documented that managerial skills measured through employee surveys predict various aspects of managerial performance (Hoffman and Tadelis, 2021; Asuyama and Owan, 2024), our measure of supervisor satisfaction may similarly provide insight into how supervisors evaluate their subordinates.

Table 4 shows the regression result. We find that supervisors who entered the firm more recently tend to be significantly more accurate evaluators. While this may reflect improved training or evaluation practices for newer cohorts, it is also conceivable that age proximity between supervisors and subordinates facilitates better understanding and communication, particularly given that our analysis focuses on lower-ranked and generally younger employees. Drinking habit is also positively associated with evaluation accuracy, and the estimate is statistically significant. This result may reflect that supervisors who participate in after-work socializing have more informal opportunities to learn about their subordinates' performance. It may also capture broader social skills, such as extroversion or interpersonal engagement,

[14] See Appendix I for more detail of the Employee Engagement Survey.

which could help supervisors form more nuanced impressions of employee contributions.

The coefficient on the estimated supervisor fixed effect is negative and statistically significant. One plausible interpretation is that supervisors who consistently give high evaluation scores have less room to distinguish between employees, because ratings are capped at the top of the scale. In such cases, the low dispersion in scores may result from mechanical limitations rather than an actual inability to assess performance differences, leading to lower measured accuracy. The other characteristics, such as gender, high-flyer status, smoking and exercise habits, and average boss satisfaction score, are not significantly associated with evaluation accuracy.

Overall, the result shows that some supervisor traits are associated with differences in evaluation accuracy. In particular, supervisors who joined the company more recently and those who report drinking habits tend to be more accurate. However, the other characteristics show no clear relationship, and a large part of the variation in accuracy remains unexplained. This suggests that there are likely other important factors not captured in our data that contribute to how accurately supervisors evaluate their subordinates. The difficulty of explaining the supervisor-specific accuracy with observable covariates echoes the findings of Metcalfe et al. (2023), who similarly document that much of the variation in managerial quality is hard to be explained by the available observables in their data.

Next, we turn to examining whether employees are more likely to be promoted when they report to supervisors who possess these three characteristics found to be associated with supervisor accuracy: recent hire, drinking, and leniency. We investigate whether the correlations between supervisor accuracy and characteristics translate into employee promotions. We use the same specification as in Equation (7) but with $D_{i,t}$ indicating a transition to a supervisor who possesses the characteristics under consideration. We separately estimate the model for each characteristic.

Figure 9 presents the effect of reporting to a supervisor who possesses a given characteristic on promotion probability over time. We examine three traits: being a drinker (panel (a)), being younger (panel (b)), and being lenient in evaluations (panel (c)). The estimates are reported relative to the year before the supervisor switch, which serves as the omitted baseline. The results suggest that drinking and being young may translate to more employee promotions. Reporting to a drinking supervisor is linked to a statistically significant increase in promotion probability in period 3. Also, employees who transition to a younger supervisor experience a marginally significant boost in promotion probability in period 2. While the overall patterns are not uniformly strong across all traits and time periods, these findings indicate that supervisor characteristics can matter for career advancement in specific contexts and time horizons.

4.5.2 Proximity Advantage in Promotions

The previous sections have documented substantial heterogeneity in supervisor evaluation accuracy and explored how this variation relates to observable characteristics such as entry cohort, fixed rater tendencies, and managerial trajectories. While some factors like recent entry and drinking have shown significant associations, other potential indicators of managerial quality, such as promotion speed or subordinate satisfaction, are not predictive of evaluation accuracy. This suggests that conventional professional attributes alone may not fully account for differences in supervisors' ability to assess subordinates accurately.

These findings raise the possibility that less formal, interpersonal dynamics may also shape evaluation accuracy. In particular, certain types of proximity or shared traits between supervisors and subordinates, such as age, tenure, or personal habits, may facilitate better mutual understanding and observation. For example, supervisors who are closer in age or tenure to their subordinates or who share behavioral patterns like smoking or drinking may have more opportunities for informal interaction. These interactions can provide deeper insight into employees' day-to-day effort and contributions. While such settings may not reflect formal managerial skill, they can nonetheless enhance evaluative accuracy by improving supervisors' exposure to and interpretation of employee performance.

Motivated by this possibility, we now consider three types of proximity advantage between supervisors and subordinates in terms of subordinate careers: *drinker-to-drinker, younger-to-younger*, and *smoker-to-smoker advantages*.[15] The idea is that when employees report to a supervisor who shares a particular trait or habit, they may get a leg up in promotions due to a stronger relationship or better mutual understanding, which may lead to more accurate evaluations.

Building on the empirical framework proposed by Cullen and Perez-Truglia (2023), we estimate the following event-study model:

$$\begin{aligned} y_{i,t} = & \left(\sum_{k \in \mathcal{K}} \beta_k^D D_{i,t-k} + \sum_{k \in \mathcal{K}} \beta_k^T T_{i,t-k} \right) \times (1 - B_i) \\ & + \left(\sum_{k \in \mathcal{K}} \theta_k^D D_{i,t-k} + \sum_{k \in \mathcal{K}} \theta_k^T T_{i,t-k} \right) \times B_i \\ & + \alpha_i + \tau_t + \varepsilon_{i,t}. \end{aligned} \tag{11}$$

[15] Due to the very limited number of female supervisors in our data, we are unable to examine gender-based proximity effects such as male-to-male or female-to-female matching. Prior studies such as Cullen and Perez-Truglia (2023) and Fortin et al. (2022) document that male-to-male advantages are substantial in terms of employee promotions and pay.

Here, $y_{i,t}$ is an indicator for whether employee i is promoted in year t. The event-study variable $T_{i,t-k}$ is defined as in Equation (7), while $D_{i,t-k}$ now indicates the timing when employee i is assigned to a supervisor with a specific characteristic: smoking, drinking, exercising, or being younger. The key extension in Equation (11) is the interaction with B_i, a dummy variable that equals one if subordinate i shares the characteristic in question with the supervisor. We consider each characteristic separately. A subordinate is coded as "younger" if their entry year is later than the median among all subordinates. This specification allows the dynamic treatment effects to vary based on whether the subordinate and supervisor share the same characteristic, enabling us to test for proximity advantages in promotion outcomes. We estimate the model separately for each characteristic.

Figure 10 presents event-study estimates of promotion probabilities around the time of supervisor switches, disaggregated by whether the subordinate shares a given characteristic with the incoming supervisor. Each panel corresponds to a different potential dimension of proximity: smoking (panel (a)), drinking (panel (b)), exercising (panel (c)), and being younger (pangel (d)). In each panel, the purple points represent subordinates who share the characteristic with the new supervisor, while the orange points show those who do not. Across all four dimensions, we do not observe consistent or statistically meaningful differences in promotion patterns between the proximity group and the control group. The estimates tend to overlap substantially, and no systematic divergence emerges before or after the supervisor switch. These null results suggest that proximity along these behavioral or demographic dimensions does not lead to detectable differences in promotion outcomes. This finding contrasts with the results of Cullen and Perez-Truglia (2023), who document evidence of male-to-male proximity advantage in promotions. In our setting, however, we do not find clear support for similar proximity-based advantages along the observed characteristics.

4.5.3 Employee Survey

Rather exploratorily, we examine whether supervisor accuracy is associated with various outcomes by regressing each item from the company's annual employee engagement survey on supervisor accuracy. The questionnaire items are provided in Appendix I. The regression equation is the same as Equation (8). Figure 11 shows estimates from each regression.[16] Although most estimates are not statistically significant, there is a general tendency for higher supervisor accuracy to be associated with more favorable employee responses. In particular, the estimate on *role-grade match* is statistically significant, which may suggest that accurate-rating supervisors are more effective at assigning roles that align with employees'

[16] We omit two survey items from the figure due to the large confidence intervals, which makes the figure difficult to interpret visually.

abilities. It is also notable that the estimate on *supervisor guidance* is marginally significant, which suggests that these supervisors may be better at providing clear and appropriate instructions, thereby enabling employees to perform their roles more effectively.

This pattern may suggest that employees tend to be more satisfied when their supervisors evaluate them accurately. Accurate evaluations may foster greater trust, a sense of fairness, and clearer expectations, which in turn can contribute to more positive perceptions of the work environment. Alternatively, it may reflect that supervisor accuracy is correlated with a particular managerial style that independently contributes to more positive employee perceptions.

5 Concluding Remarks

This paper has shown that variation in supervisors' ability to translate noisy signals into performance ratings has first-order implications for employee careers. We developed a simple model in which a supervisor receives a signal of her subordinate's performance and translate them into assessment scores. The model predicts that more accurate supervisors generate greater dispersion in ratings across subordinates. Using rich personnel records from a large Japanese manufacturing firm, we confirm this prediction empirically: supervisors with higher within-rater variance are more likely to promote their subordinates. We also find evidence consistent with other predictions of the model: accurate raters are more effective at distinguishing performance levels among subordinates and may contribute to the long-run success of the employees they promote, even after they move on to new supervisors. Overall, these results demonstrate that supervisor accuracy, although difficult to observe directly, plays a critical role in shaping internal labor market outcomes.

These findings have practical implications for how firms design and monitor their evaluation systems. Because accuracy shows up as greater variation in scores across subordinates, organizations can use this measure to identify supervisors who give overly similar or uniformly high ratings, which reduces the usefulness of evaluations. Simple diagnostic tools, built into existing HR dashboards, could help flag these cases and support better performance management. In addition, calibration sessions, targeted feedback, and rater training may help promote more thoughtful and accurate assessments. For example, Demeré et al. (2019) and Grabner et al. (2020) show that calibration committees can reduce evaluation bias among supervisors. It is also notable that Manthei and Sliwka (2019) and Bernstein and Li (2025) find that making performance data available to supervisors or employees can increase employee effort, further underscoring the value of performance transparency in

shaping behavior.[17]

We also find suggestive evidence that supervisors who appear to have more opportunities to socialize or stronger social skills, as indicated by their drinking habits, tend to be more accurate in their evaluations. This may reflect that interpersonal engagement or social skills help supervisors observe and understand employee performance more clearly. Firms may benefit from encouraging structured opportunities for interaction that support better evaluation, while still avoiding favoritism or unfair bias. This view aligns with prior research suggesting that social or interpersonal skills are important traits of effective managers and supervisors (Artz et al., 2017; Kuroda and Yamamoto, 2018; Hansen et al., 2021; Hoffman and Tadelis, 2021; Asuyama and Owan, 2024). At the same time, caution is warranted. Other studies have shown that social ties can foster favoritism, undermining merit-based decision-making and potentially leading to inefficient allocation of resources (Bandiera et al., 2009; Cullen and Perez-Truglia, 2023; Wang et al., 2023).

Several limitations point to promising directions for future research. First, as is typical of any insider econometric work, our analysis focuses on a single firm in one national context, so applying the framework to other sectors, countries, or organizational structures would help assess its external validity. Second, although we use quasi-random variation in supervisor assignment, stronger causal designs such as experimental interventions or institutional changes could more clearly identify the effect of supervisor accuracy. Third, as is common in research on employee evaluations, we lack objective measures of performance or ability, which are rarely available to researchers or even to firms themselves, with the notable exception documented by Altonji and Pierret (2001). Although we partially address this challenge by providing a clear theoretical framework to guide our empirical strategy, validating our measure of supervisor accuracy against objective performance metrics would offer more convincing tests.

A particularly promising direction for future research is to examine the long-term organizational and market-level consequences of improved evaluation accuracy. As Pallais (2014) demonstrates, more accurate evaluations can enhance not only employee outcomes but also overall market performance. Such work would deepen our understanding of how frontline evaluation practices shape outcomes beyond the individual level and at the organizational and market scale. It would also be valuable to explore how assessment technologies such as digital tools, algorithmic rating systems, or peer evaluations interact with or potentially substitute for supervisor judgment. These questions are left for future research.

[17] There is also experimental evidence from lab and online settings on how to design supervisor incentives or assessment structures to enhance the effectiveness of subjective employee evaluations (Ockenfels et al., 2024; Kusterer and Sliwka, 2025).

References

Agarwal, S., Qian, W., Reeb, D. M., and Sing, T. F. (2016). Playing the boys game: Golf buddies and board diversity. *American Economic Review*, 106(5):272–276.

Altonji, J. G. and Pierret, C. R. (2001). Employer learning and statistical discrimination. *The Quarterly Journal of Economics*, 116(1):313–350.

Artz, B. M., Goodall, A. H., and Oswald, A. J. (2017). Boss competence and worker well-being. *IlR Review*, 70(2):419–450.

Asuyama, Y. and Owan, H. (2024). People management skills, senior leadership skills and the peter principle. Technical report, RIETI Discussion Paper.

Baker, G., Gibbons, R., and Murphy, K. J. (1994). Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics*, 109(4):1125–1156. Publisher: Oxford University Press.

Bandiera, O., Barankay, I., and Rasul, I. (2009). Social connections and incentives in the workplace: Evidence from personnel data. *Econometrica*, 77(4):1047–1094.

Benson, A., Li, D., and Shue, K. (2024). Potential and the gender promotion gap. *SSRN*.

Bernstein, E. and Li, S. X. (2025). The performance effects of giving front-line employees direct access to performance data and thereby limiting the supervisor's feedback-intermediation role: evidence from a field experiment. *Management Science*.

Biggerstaff, L. E., Campbell, J. T., and Goldie, B. A. (2024). Hitting the "grass ceiling": Golfing CEOs, exclusionary schema, and career outcomes for female executives. *Journal of Management*, 50(5):1502–1535.

Bushman, R. M., Indjejikian, R. J., and Smith, A. (1996). CEO compensation: The role of individual performance evaluation. *Journal of Accounting and Economics*, 21(2):161–193.

Cullen, Z. and Perez-Truglia, R. (2023). The old boys' club: Schmoozing and the gender gap. *American Economic Review*, 113(7):1703–1740.

Demeré, B. W., Sedatole, K. L., and Woods, A. (2019). The role of calibration committees in subjective performance evaluation systems. *Management Science*, 65(4):1562–1585.

Diaz, B. S., Nazarrett, A. N., Ramirez, J., Sadun, R., and Tamayo, J. A. (2025). Training within firms. NBER Working Paper 33670.

Elvira, M. and Town, R. (2001). The effects of race and worker productivity on performance evaluations. *Industrial Relations: A Journal of Economy and Society*, 40(4):571–590. Publisher: John Wiley & Sons, Ltd.

Fortin, N. M., Markevych, M., and Rehavi, M. (2022). Closing the gender pay gap in the us federal service: the role of new managers. Technical report, Working Paper.

Frederiksen, A., Kahn, L. B., and Lange, F. (2020). Supervisors and performance management systems. *Journal of Political Economy*, 128(6):2123–2187.

Frederiksen, A., Lange, F., and Kriechel, B. (2017). Subjective performance evaluations and employee careers. *Journal of Economic Behavior & Organization*, 134:408–429.

Friebel, G., Heinz, M., Hoffman, M., Kretschmer, T., and Zubanov, N. (2024). Is this really kneaded? identifying and eliminating potentially harmful forms of workplace control. Technical report, ECONtribute Discussion Paper.

Friebel, G., Heinz, M., and Zubanov, N. (2022). Middle managers, personnel turnover, and performance: A long-term field experiment in a retail chain. *Management Science*, 68(1):211–229.

Fuchs, W. (2007). Contracting with repeated moral hazard and private evaluations. *The American Economic Review*, 97(4):1432–1448. Publisher: American Economic Association.

Gibbs, M., Merchant, K. A., Van der Stede, W. A., and Vargus, M. E. (2004). Determinants and effects of subjectivity in incentives. *The Accounting Review*, 79(2):409–436. Publisher: American Accounting Association.

Grabner, I., Künneke, J., and Moers, F. (2020). How calibration committees can mitigate performance evaluation bias: An analysis of implicit incentives. *The Accounting Review*, 95(6):213–233.

Greene, W. H. (2003). *Econometric Analysis*. Prentice Hall, Upper Saddle River, NJ, 5 edition.

Haegele, I. (2024). Talent hoarding in organizations. Working paper. Available at https://arxiv.org/pdf/2206.15098.

Hansen, S., Ramdas, T., Sadun, R., and Fuller, J. (2021). The demand for executive skills. Technical report, National Bureau of Economic Research.

Hayes, R. M. and Schaefer, S. (2000). Implicit contracts and the explanatory power of top executive compensation for future performance. *The RAND Journal of Economics*, 31(2):273–293. Publisher: [RAND Corporation, Wiley].

Hoffman, M., Kahn, L. B., and Li, D. (2018). Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800.

Hoffman, M. and Tadelis, S. (2021). People management skills, employee attrition, and manager rewards: An empirical analysis. *Journal of Political Economy*, 129(1):243–285.

Holmstrom, B. and Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization*, 7:24–52. Publisher: Oxford University Press.

Izumi, Y., Shigeoka, H., and Yagasaki, M. (2024). Golfing CEOs. *Labour Economics*, 91:102639.

Kampkötter, P. and Sliwka, D. (2018). More dispersion, higher bonuses? on differentiation in subjective performance evaluations. *Journal of Labor Economics*, 36(2):511–549.

Kawaguchi, D., Owan, H., and Takahashi, K. (2016). Biases in subjective performance evaluation. RIETI Discussion Paper Series 16-E-059.

Kawata, Y. and Owan, H. (2022). Peer effects on job satisfaction from exposure to elderly workers. *Journal of the Japanese and International Economies*, 63:101183.

Kuroda, S. and Yamamoto, I. (2018). Good boss, bad boss, workers' mental health and productivity: Evidence from japan. *Japan and the World Economy*, 48:106–118.

Kusterer, D. J. and Sliwka, D. (2025). Social preferences and the informativeness of subjective performance evaluations. *Management Science*, 71(5):4028–4048.

Lazear, E. P. (2004). The peter principle: A theory of decline. *Journal of Political Economy*, 112(S1):S141–S163.

Lazear, E. P., Shaw, K. L., and Stanton, C. T. (2015). The value of bosses. *Journal of Labor Economics*, 33(4):823–861.

Levin, J. (2003). Relational incentive contracts. *The American Economic Review*, 93(3):835–857. Publisher: American Economic Association.

Macdonald, D. C., Montonen, J., and Nix, E. E. (2025). The impacts of romantic relationships with the boss. NBER Working Paper 31192.

MacLeod, W. B. (2003). Optimal contracting with subjective evaluation. *The American Economic Review*, 93(1):216–240. Publisher: American Economic Association.

Manthei, K. and Sliwka, D. (2019). Multitasking and subjective performance evaluations: Theory and evidence from a field experiment in a bank. *Management Science*, 65(12):5861–5883.

Medoff, J. L. and Abraham, K. G. (1980). Experience, performance, and earnings. *The Quarterly Journal of Economics*, 95(4):703–736. Publisher: Oxford University Press.

Metcalfe, R. D., Sollaci, A. B., and Syverson, C. (2023). Managers and productivity in retail. NBER Working Paper 31192.

Minni, V. (2023). Making the invisible hand visible: Managers and the allocation of workers to jobs. CEP Discussion Paper CEPDP1948.

Morita, K. (2025). Developmental role of negative feedback. *Available at SSRN 3431572*.

Ockenfels, A., Sliwka, D., and Werner, P. (2024). Multi-rater performance evaluations and incentives.

Pallais, A. (2014). Inefficient hiring in entry-level labor markets. *American Economic Review*, 104(11):3565–3599.

Prendergast, C. and Topel, R. (1993). Discretion and bias in performance evaluation. *European Economic Review*, 37(2):355–365.

Prendergast, C. and Topel, R. H. (1996). Favoritism in organizations. *Journal of Political Economy*, 104(5):958–978. Publisher: University of Chicago Press.

Roberts, J. and Shaw, K. L. (2022). Managers and the management of organizations. NBER Working Paper 30730.

Shukla, S. (2025). Making the elite: Class discrimination at multinationals.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

Takahashi, S., Owan, H., Tsuru, T., and Uehara, K. (2021). Multitasking incentives and the informative value of subjective performance evaluations. *ILR Review*, 74(2):511–543. Publisher: SAGE Publications Inc.

Thiele, V. (2013). Subjective performance evaluations, collusion, and organizational design. *Journal of Law, Economics, & Organization*, 29(1):35–59. Publisher: Oxford University Press.

Wang, J., Huang, C., Xu, L., and Zhang, J. (2023). Drinking into friends: Alcohol drinking culture and CEO social connections. *Journal of Economic Behavior & Organization*, 212:982–995.

Wu, H. X. and Liu, S. X. (2020). The trade-offs of letting local managers make hiring decisions. *Available at SSRN 3500302*.

Figures and Tables

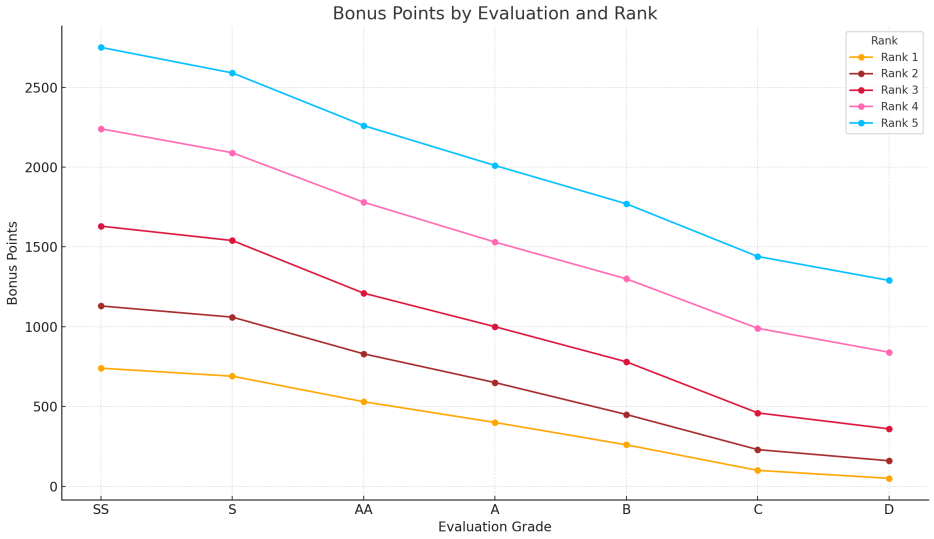


Figure 1: Bonus Points By Evaluation And Pay Grade

Notes: The figure shows the distribution of bonus points awarded to employees by evaluation grade, with colors indicating their current pay grade. Higher evaluation grades are associated with more bonus points, and within each evaluation grade, employees at higher pay grades tend to receive slightly more bonus points on average. This reflects the firm's bonus policy, which links rewards to both performance and position level.

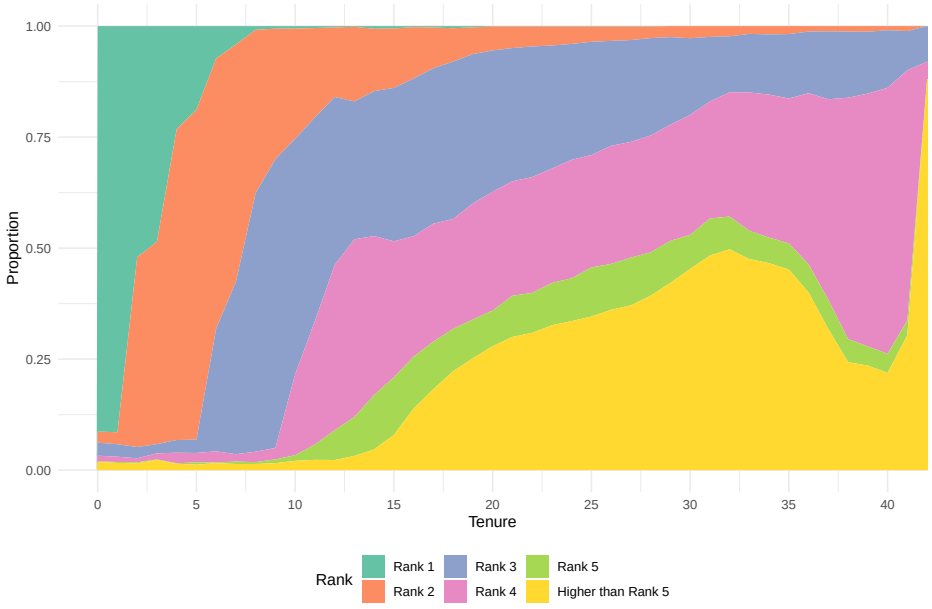


Figure 2: The Distribution of Employee Pay Grades by Tenure

Notes: The figure illustrates how the distribution of pay grades evolves with employee tenure. While nearly all employees begin at grade Rank 1, most are promoted to Rank 2 and Rank 3 within the first five years. After that, promotion timing becomes more heterogeneous, with some employees advancing to higher grades such as Rank 4, Rank 5, and above, while others remain in lower grades even after long tenures.

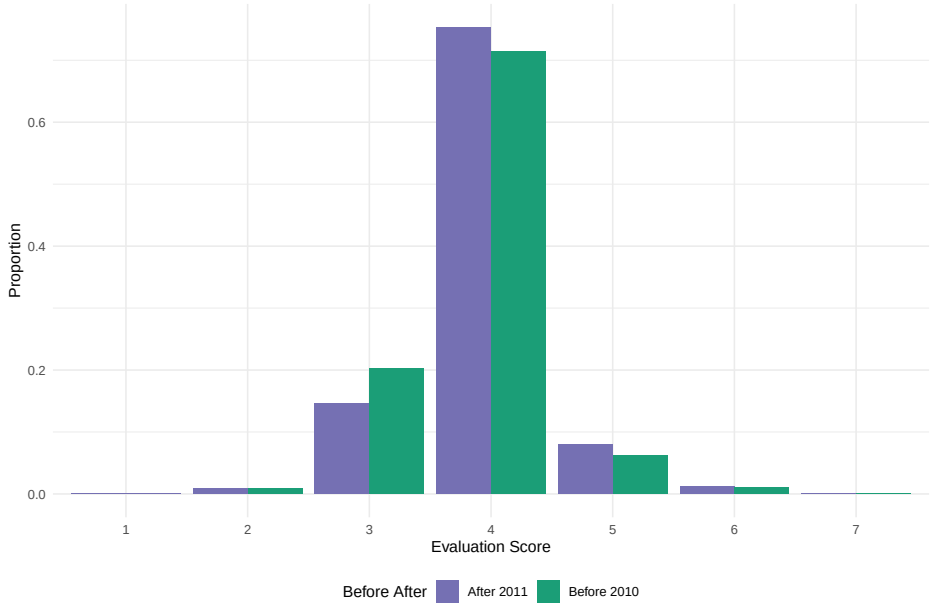


Figure 3: The Distribution of Employee Evaluation Score

Notes: The figure compares the distributions of ratings before and after the labeling change. The green bars show scores before 2010. The blue bars show scores after 2011. See Section 3.2 for details about the change in the firm's labeling system.

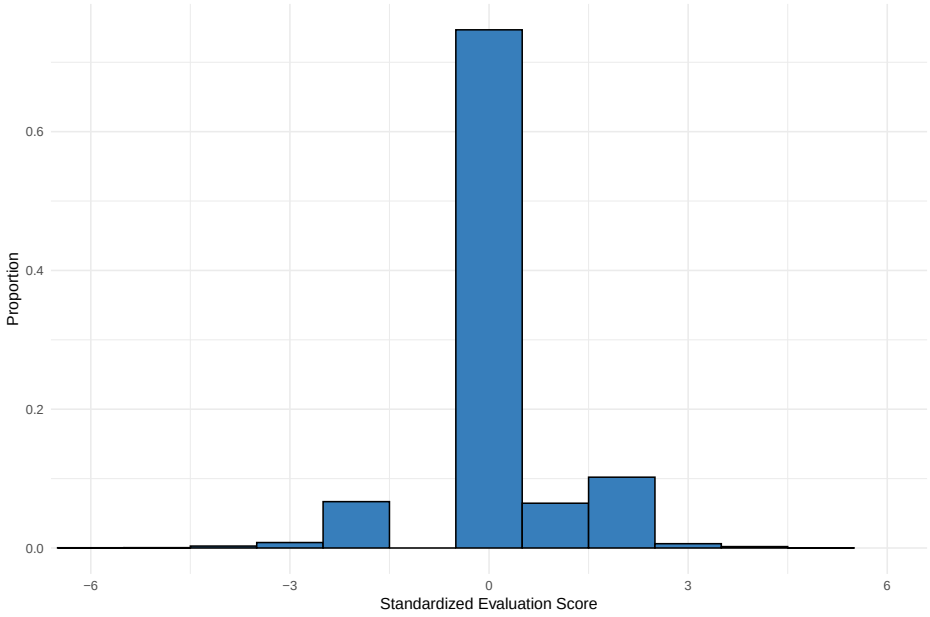


Figure 4: The Distribution of Nomarlized Employee Evaluation Score

Notes: The figure shows the distribution of standardized employee performance evaluation scores. Each score is normalized by subtracting the fiscal-year mean and dividing by the fiscal-year standard deviation. See Section 3.2 for details about the scoring-system reform.

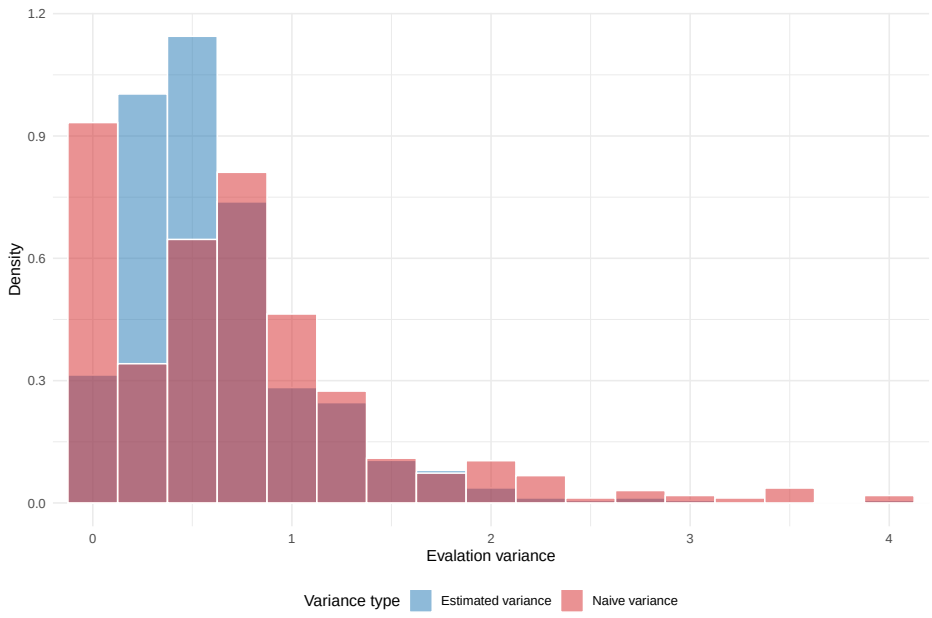


Figure 5: The Distribution of Supervisor Evaluation Variance

Notes: The figure compares the distributions of supervisor-specific evaluation variances across two measures. The red histogram shows the naive variance calculated directly from raw assessment scores. The blue histogram shows the estimated variance based on residualized scores after removing employee, supervisor, division, and year fixed effects. See Section 4.1 for details about the measurement.

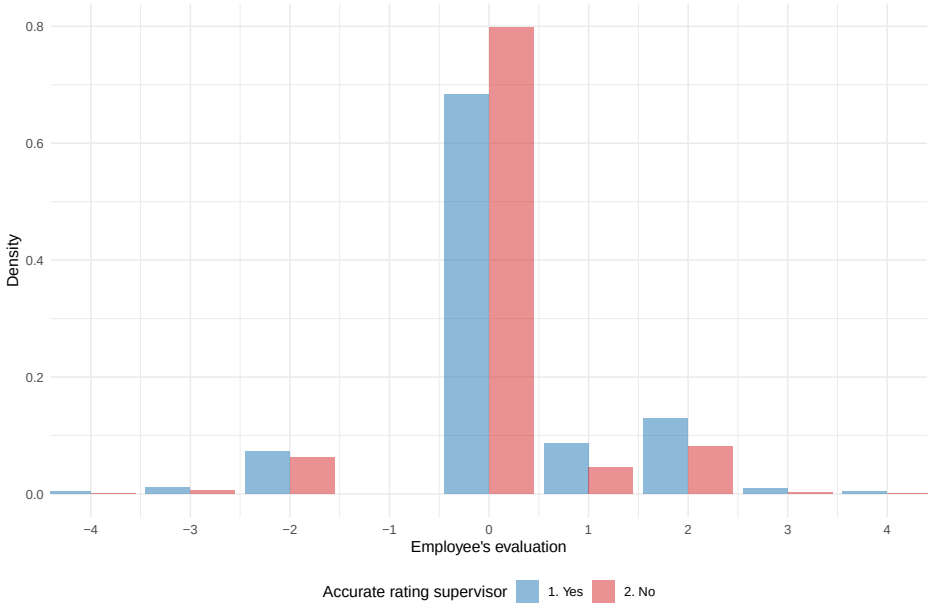


Figure 6: The Distributions of Employee Evaluations by Accurate and Non-Accurate Raters

Notes: The figure shows the distribution of employee evaluation scores, standardized within each year, separately for employee reporting to accurate-rating supervisors (blue) and those assigned to less accurate-rating supervisors (red). Supervisors whose estimated evaluation variance is above the median are classified as accurate raters. While both groups are centered around zero, evaluations from accurate raters display greater dispersion. This pattern aligns with the model's prediction that higher evaluation accuracy corresponds to greater variance in performance assessments.

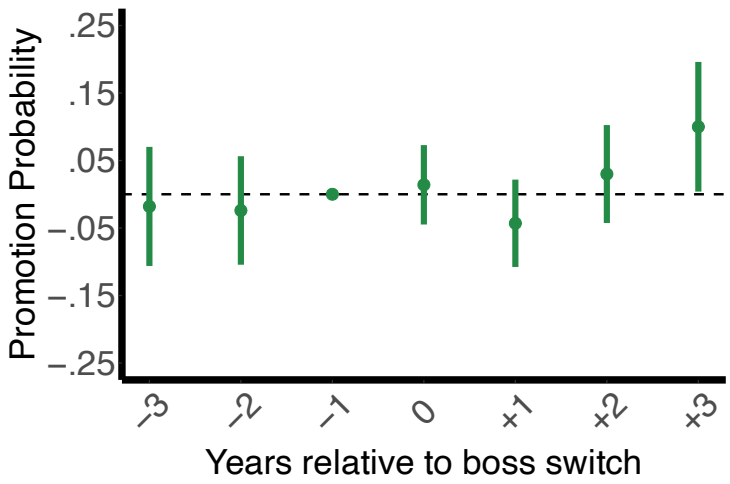


Figure 7: The Dynamic Effects of Accurate Raters on Employee Promotions

Notes: See Section 4.2 for details about the model specification. The figure presents the estimation results from the event-study model. Each dot represents the estimated effect of reporting to an accurate-rating supervisor in each event period (estimate of β_k^D). The vertical segments represent the 95 percent confidence intervals. The standard errors are clustered at the employee and supervisor levels. Period 0 is the exact year when employees start reporting to new supervisors. Period -1 is omitted as the baseline in estimation, so the coefficient for period -1 is zero by construction.

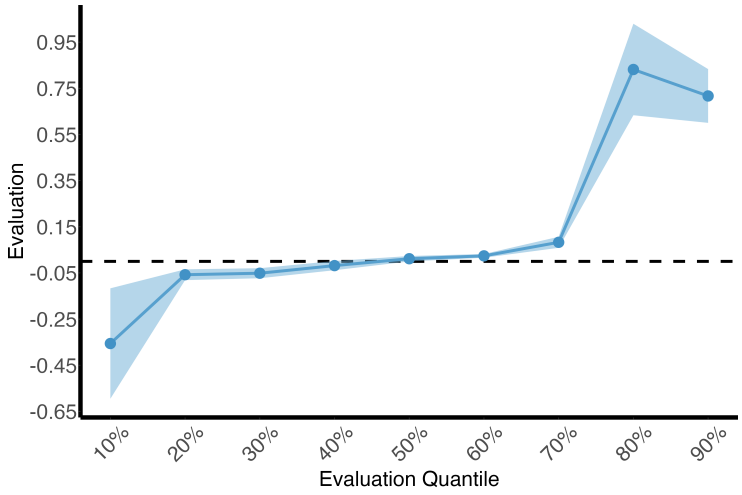


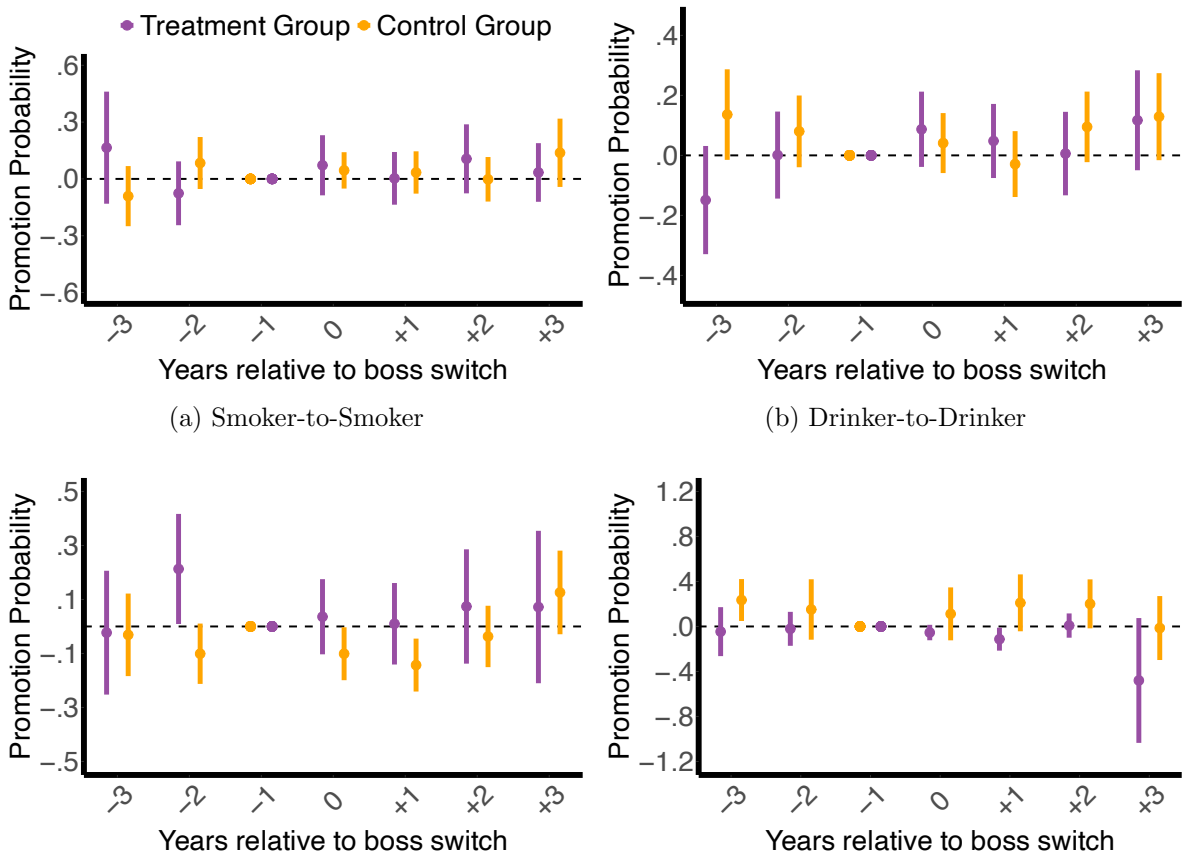
Figure 8: Quantile Regression of Performance Evaluation on Supervisor Accuracy

Notes: See Section 4.3 for details on the model specification. The figure presents the results from a quantile regression of employee evaluations on supervisor accuracy. Each point in the figure corresponds to the estimated coefficient on supervisor accuracy at a given evaluation quantile (i.e., estimates of ρ_q), and the vertical lines represent 95 percent confidence intervals.

(a) Drinking Supervisor (b) Younger Supervisor (c) Lenient Supervisor

Figure 9: The Effect of Supervisors with Certain Characteristics

Notes: See Section 4.5.1 for details about the model specification. The figure presents the estimation results from the event-study model. Each dot represents the estimated effect of reporting to a supervisor who possess a certain characteristic in each event period (estimate of β_s^D). The vertical segments represent the 95 percent confidence intervals. The standard errors are clustered at the employee and supervisor levels. Period 0 is the exact year when employees start reporting to new supervisors. Period -1 is omitted as the baseline in estimation, so the coefficient for period -1 is zero by construction.



(a) Smoker-to-Smoker

(b) Drinker-to-Drinker

(c) Exerciser-to-Exerciser

(d) Younger-to-Younger

Figure 10: Proximity Advantage in Promotions

Notes: See Section 4.5.2 for details about the model specification. The figure presents the estimation results from the event-study model. Each dot represents the estimated effect of reporting to a supervisor who possesses a certain characteristic in each event period (estimates of θ_k^D and β_k^D). The purple bar corresponds to employees in a treatment group, and the yellow bar corresponds to employees in a control group. The vertical segments represent the 95 percent confidence intervals. The standard errors are clustered at the employee and supervisor levels. Period 0 is the exact year when employees start reporting to new supervisors. Period -1 is omitted as the baseline in estimation, so the coefficient for period -1 is zero by construction.

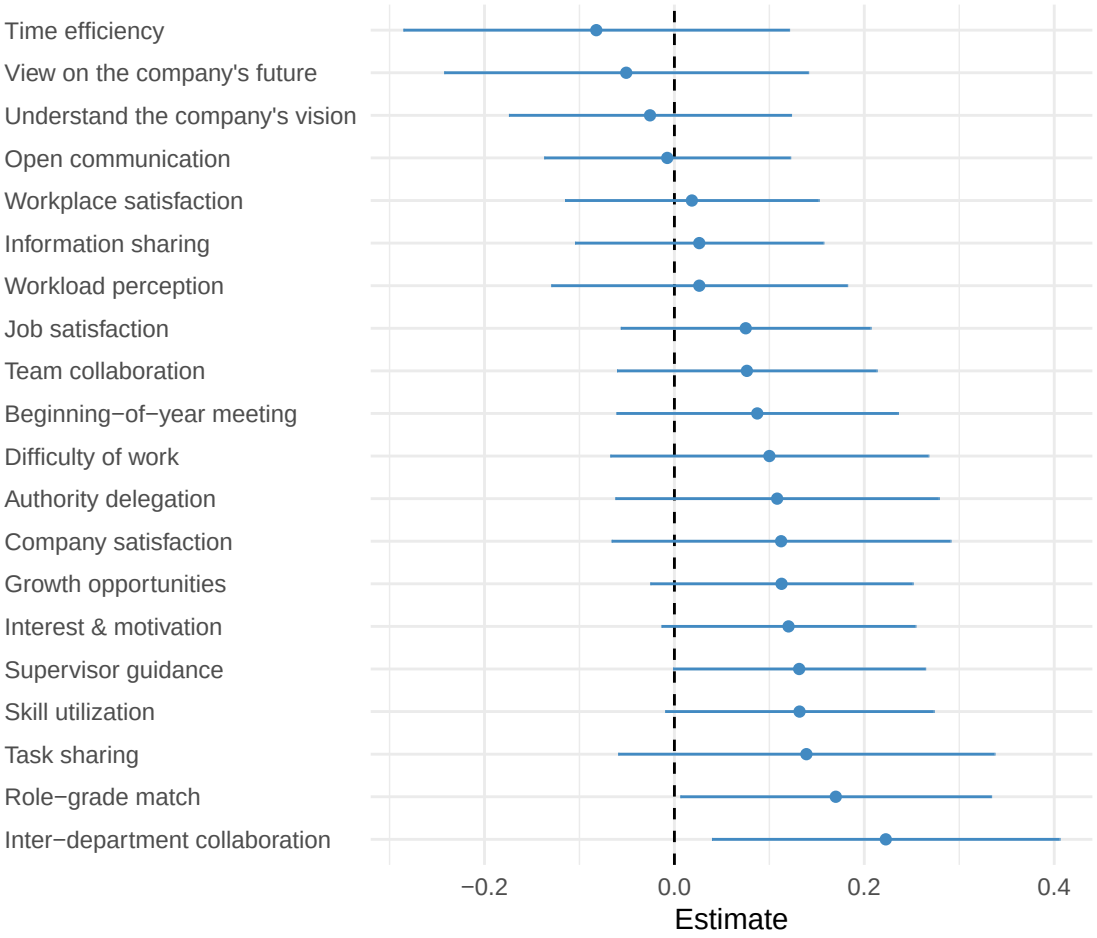


Figure 11: Employee Engagement Survey Responses and Supervisor Accuracy

Notes: See Section 4.5.3 for details about the model specification. The figure presents the estimation results from regressions of employee engagement survey responses on supervisor accuracy. Each dot represents the estimate obtained by regressing each employee engagement survey response on supervisor accuracy. The horizontal bars indicate 95 percent confidence intervals. All regressions control for employee and year fixed effects, as well as the squares of age and tenure. Survey responses are coded such that higher values indicate more favorable perceptions.

Table 1: Summary Statistics of Estimated Supervisor Accuracy

Supervisor Accuracy:	Mean	Median	Variance	S.D.	10th	25th	75th	90th
$\widehat{\sigma}_{s(i,t)}$	0.71	0.70	0.13	0.35	0.24	0.51	0.90	1.14

Notes: The table reports the summary statistics of estimated supervisor accuracy in subjective evaluations. Supervisor accuracy is measured in two ways: the standard deviation (S.D.) and the variance (Var.) of residuals from Equation (5). Higher values indicate higher accuracy. The mean, median, and selected percentiles are reported across supervisors.

Table 2: Supervisor Evaluation Accuracy and Employee Promotions

Outcome: Promotion	(1)	(2)	(3)	(4)	(5)
Supervisor Accuracy ($\widehat{\sigma}_{s(i,t)}$)	0.0561***	0.0917**	0.0398**	0.0838**	0.0808**
	(0.0187)	(0.0364)	(0.0183)	(0.0362)	(0.0362)
Employee FEs	No	Yes	No	Yes	Yes
Year FEs	No	No	Yes	Yes	Yes
Age^2 & Tenure^2	No	No	No	No	Yes
Observations	6,512	6,512	6,512	6,512	6,512

Notes: See Section 4.2.2 for details on the model specification. The table shows the estimation results of regressions of promotions on supervisor accuracy. The outcome is a binary indicator for whether the employee is promoted in a given year. The key independent variable is the square root of the estimated variance of supervisor ratings. Columns vary in included fixed effects. The numbers in parentheses are standard errors clustered at the employee and supervisor levels. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 3: Quasi-Poisson Count Model Estimation Result

Outcome	# of Promotions After 2nd Boss Switch	
	(1)	(2)
Variable	Coefficients	Average Marginal Effects
Accurate Rater	0.0759*	0.0889*
	(0.0425)	(0.0499)
Age & Age^2	Yes	Yes
Tenure & Tenure^2	Yes	Yes
Observations	1,075	1,075

Notes: The table shows the estimation result from the quasi-Poisson regression model. The dependent variable is the number of promotions after the second supervisor switch. The key independent variable is the square root of the estimated variance of supervisor ratings. The numbers in parentheses are standard errors clustered at the employee and supervisor levels. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 4: Linear Regression of Supervisor Accuracy on Supervisor Characteristics

Dependent Variable:	Supervisor Accuracy ($\widehat{\sigma}_s$)	
Entry Year	0.3624^{***}	(0.0756)
High-Flyer	0.0100	(0.0646)
Drinking Habit	0.0899^{**}	(0.0436)
Smoking Habit	0.0766	(0.0476)
Exercise Habit	0.0219	(0.0481)
Supervisor Leniency	-0.4551^{***}	(0.1358)
Average Boss Satisfaction Score	-0.0668	(0.1145)
Female	-0.2071	(0.1815)

Notes: The table shows estimates from the linear regression of supervisor accuracy (measured by the dispersion of evaluation scores) on observable supervisor characteristics. The regression includes demographic variables (gender, entry year), behavioral traits (drinking, smoking, and exercise habits), a "high-flyer" dummy, supervisor leniency (supervisor fixed effects from equation (4)), and average boss satisfaction scores. The numbers in parentheses are standard errors clustered at the employee and supervisor levels. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Appendix A Formalizing the Supervisor's Problem

Here, we provide a more formal treatment of the supervisor's problem. Assume that the supervisor exerts effort $e \in \mathbb{R}_+$, and this reduces the signal noise such that $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2/e)$. The supervisor incurs cost of effort $c(e)$, where $c(0) = 0, c' > 0, c'' > 0$, and then he receives a signal and submits his rating score. The employer observes the evaluation submitted by the supervisor. She also observes the true performance of the employee with the probability $p((r-q)^2)$, where r is the submitted rating score, and we assume $p \in [0,1], p(0) = 0, p' > 0, p'' > 0$. The supervisor's expected payoff is given by

$$U = W - P \times \mathsf{E}\left[\, p\left((r-q)^2\right)\,\right] - \kappa\, c(e),$$

where W and P are the fixed wage and fine, and $\kappa > 0$ governs the cost parameter that captures heterogeneity in supervisor accuracy. Note that the optimal rating for the supervisor r^* is made after his effort is sunk and the signal is received, so the optimal rating is characterized by

$$r^* = \arg\max_r \mathsf{E}\left[\, -P \times \mathsf{E}\left[\, p\left((r-q)^2\right)\,\right] \mid \widehat{q}\,\right].$$

Since p is an increasing function, r^* almost coincides with Equation (2). Namely,

$$r^*(e) = \frac{\sigma_\varepsilon^2/e}{\sigma_q^2 + \sigma_\varepsilon^2/e}\overline{q} + \frac{\sigma_q^2}{\sigma_q^2 + \sigma_\varepsilon^2/e}\widehat{q}.$$

The first-order condition for the supervisor is

$$\kappa\, c'(e) = -P \cdot h'(e),$$

where $h(e) = \mathsf{E}\left[\, p\left((r^*(e) - q)^2\right)\,\right]$. Note that $r^*(e) - q = \frac{\sigma_q^2}{\sigma_q^2+\sigma_\varepsilon^2/e}\varepsilon - \frac{\sigma_\varepsilon^2/e}{\sigma_q^2+\sigma_\varepsilon^2/e}q + \frac{\sigma_\varepsilon^2/e}{\sigma_a^2+\sigma_\varepsilon^2/e}\overline{q}$. So, $r^*(e) - q \sim \mathcal{N}\left(0, \frac{\sigma_q^2\sigma_\varepsilon^2/e}{\sigma_q^2+\sigma_\varepsilon^2/e}\right)$. Let $V(e) = \frac{\sigma_q^2\sigma_\varepsilon^2/e}{\sigma_q^2+\sigma_\varepsilon^2/e}$. Suppose $Z \sim \mathcal{N}(0,1)$, then $r^*(e) - q \overset{\mathsf{d}}{=} \sqrt{V(e)}\, Z$, which implies $(r^*(e)-q)^2 \overset{\mathsf{d}}{=} V(e)\, Z^2$. Now we can write $h(e) = P \cdot \mathsf{E}\left[\, p\left(V(e)Z^2\right)\,\right]$. By Leibniz' rule, we obtain

$$h'(e) = -P\frac{\sigma_q^4\sigma_\varepsilon^2}{(\sigma_q^2 e + \sigma_\varepsilon^2)^2}\mathsf{E}\left[\, Z^2 p'\left(V\left(e\right)Z^2\right)\,\right].$$

Clearly, all the terms are positive. Because $c'(e)$ is strictly increasing and $-h'(e)$ is decreasing in e, their graphs intersect at most once, and the optimal effort e^* is uniquely pinned down. The first-order condition implies that $\partial e^*/\partial\kappa < 0$, i.e., supervisors with a higher cost parameter κ choose lower effort. Reduced effort increases the posterior variance $V(e^*)$, generating noisier performance ratings. This strict and monotone mapping from κ to $V(e^*)$ justifies the parametric assumption we impose on the signal variances σ_A^2 and σ_N^2 in Section 2.

Let $\underline{W}$ denote the reservation utility of the supervisor. The employer solves the following

cost minimization problem:

$$\max_{W,P} P \times \mathsf{E}\left[\, p((r-q)^2) \,\right] - \gamma \mathsf{E}[\,(r-q)^2\,] - W \quad \text{s.t.} \quad U \geq \underline{W},$$
$$r = r^*(e),$$
$$e = e^*.$$

Solving the firm's optimization problem is straightforward but not essential for our purpose, so we omit it here. What should be highlighted here is that once the supervisor's participation constraint is met and the incentive pair (W, P) is chosen optimally, the remaining problem reduces to the supervisor's problem we consider in Section 2.

Appendix B Alternative Models of Supervisor Decision

It is possible that heterogeneity in supervisors' assessment dispersion is driven by mechanisms other than their ability to discern performance. In this appendix, we explore alternative modeling strategies and demonstrate that our main message remains mostly unchanged, with particular attention to leniency bias and talent hoarding.

B.1 Uniform Leniency

Supervisors may prefer to give favorable evaluations. They may gain psychological satisfaction from doing so, and they may also have incentives to motivate subordinates by giving high scores or to avoid demoralizing them by assigning low scores. So, we modify the supervisor's problem specified in Equation (1) as follows:

$$\max_{r} \Big[\mathsf{E}[\, -(r-q)^2 \mid \widehat{q}\,] + \lambda(r-\bar{q}) \Big], \tag{B1}$$

where $\lambda > 0$ parameterizes the supervisor's preference for giving favorable evaluations. It is straightforward to see that the optimal rating r^*_{UL} is given by

$$r^*_{\mathrm{UL}} = r^* + \frac{\lambda}{2}, \tag{B2}$$

where r^* is specified in Equation (2). Thus, we obtain

$$\mathsf{Var}[\, r^*_{\mathrm{UL}}\,] = \mathsf{Var}[\, r^*\,], \tag{B3}$$

which implies that our measure of accuracy is theoretically independent of leniency, theoretically speaking. Empirically, however, greater leniency may still compress evaluations because scores are discrete and top-coded. We partially address this issue by residualizing ratings on the leniency component when calculating supervisor-specific score dispersions, as in Equation (5). As long as our control for the leniency component is sufficient, our theoretical predictions continue to hold.

B.2 Inequality Aversion

Supervisors may be reluctant to differentiate among employees. This may arise from a psychological cost associated with giving unequal evaluations to subordinates, or from incentive concerns, such as the desire to avoid demoralizing them or inducing uncooperative behavior. So, we modify the supervisor's problem specified in Equation (1) as follows:

$$\max_{r} \Big[\mathsf{E}[\, -(r-q)^2 \mid \widehat{q}\,] - \lambda \mathsf{Var}[\, r\,] \Big], \tag{B4}$$

where $\lambda > 0$ parameterizes the supervisor's aversion to unequalized evaluations. Let r^*_{IA} denote the optimal rating, then we can show that

$$r^*_{\text{IA}} = \frac{\lambda}{1+\lambda}\bar{q} + \frac{1}{1+\lambda}r^*, \tag{B5}$$

where r^* is specified in Equation (2) and $\kappa = \frac{\sigma_q^2}{\sigma_q^2+\sigma_s^2}$. Note that

$$\mathsf{Var}[\, r^*_{\text{IA}}\,] = \frac{1}{(1+\lambda)^2}\mathsf{Var}[\, r^*\,] = \frac{\kappa}{(1+\lambda)^2}, \tag{B6}$$

where $\kappa = \frac{\sigma_q^4}{\sigma_q^2+\sigma_s^2}$. The variance of r^*_{IA} is decreasing in λ. This result implies that accuracy and leniency cannot be separately identified from realized dispersions. However, note that we can reinterpret $\frac{\kappa}{(1+\lambda)^2}$ (or its square root) as a measure of the supervisor's (comprehensive) ability to distinguish subordinate performance. Our theoretical predictions continue to hold we redefine accuracy such that supervisors with higher $\frac{\kappa}{(1+\lambda)^2}$ are classified as accurate, and those with lower values as non-accurate.

B.3 Negative Feedback Aversion

Supervisors may be reluctant to give negative evaluations to employees. It may be because they simply incur psychological cost to give negative evaluations to subordinates. Or it can be because of real incentive issue such that they do not want to demoralize subordinates. So we modify the supervisor's problem specified in Equation (1) as follows:

$$\max_r \Big[\mathsf{E}[\,-(r-q)^2 \mid \widehat{q}\,] - \lambda \max\{0, \bar{q} - r\}\Big]. \tag{B7}$$

where $\lambda > 0$ parameterizes the supervisor's aversion to negative evaluations. Let r^*_{NFA} denote the optimal rating, then we can show that

$$r^*_{\text{NFA}} = \begin{cases} r^* & \text{if} \quad r^* \geq \bar{q}, \\ \bar{q} & \text{if} \quad \bar{q} - \frac{\lambda}{2} < r^* < \bar{q}, \\ r^* + \frac{\lambda}{2} & \text{if} \quad r^* \leq \bar{q} - \frac{\lambda}{2}, \end{cases} \tag{B8}$$

where r^* is specified in Equation (2). The variance of r^*_{NFA} is rather complicated, but Equation (B8) is clear enough to see that $\mathsf{Var}[\, r^*_{\text{NFA}}\,]$ is decreasing in λ. This result implies that accuracy and leniency cannot be separately identified from realized dispersions. However, again, we can interpret the pair of (σ_s^2, λ) as the comprehensive parameter that governs supervisor accuracy in evaluation, and define supervisors with higher dispersion as accurate raters. If σ_s^2 and λ are positively correlated, then (σ_s^2, λ) is well-summarized by $\mathsf{Var}[\, r^*_{\text{NFA}}\,]$. If σ_s^2 and λ are negatively correlated, then (σ_s^2, λ) is noisily summarized by $\mathsf{Var}[\, r^*_{\text{NFA}}\,]$, and our estimates will be biased towards zero. In this case, we can think of our estimates as lower bounds.

B.4 Talent Hoarding

Supervisors may have an incentive to hoard talent. Namely, they may try to keep good subordinates and to let go bad subordinates. So, we modify the supervisor's problem specified in Equation (1) as follows:

$$\max_{r} \mathsf{E}[\,-(r-q)^2 + \eta(t-r)(q-t) \mid \widehat{q}\,], \tag{B9}$$

where $\eta > 0$ parameterizes the intensity of how much the supervisor is incentivized to hoard talent. The intuition of this specification is that we assume that the supervisor wants to keep employees whose performance is good enough to be promoted and let go those whose performance is not good enough. We can show that Let r^*_{TH} denote the optimal rating, then we can show that

$$r^*_{\text{TH}} = \left(1 - \frac{\eta}{2}\right) r^* - \frac{\eta}{2} t, \tag{B10}$$

where r^* is specified in Equation (2). Note that

$$\mathsf{Var}[\,r^*_{\text{TH}}\,] = \left(1 - \frac{\eta}{2}\right)^2 \mathsf{Var}[\,r^*\,] = \left(1 - \frac{\eta}{2}\right)^2 \kappa, \tag{B11}$$

where $\kappa = \frac{\sigma_q^4}{\sigma_q^2 + \sigma_s^2}$. This result implies that accuracy and leniency cannot be separately identified from realized dispersion.

Notice that the variance of r^*_{TH} is decreasing in η over $\eta \in (0, 2]$. When all supervisors are characterized by $\eta \in (0, 2]$, we can reinterpret $\left(1 - \frac{\eta}{2}\right)^2 \kappa$ (or its square root) as a measure of the supervisor's (comprehensive) ability to distinguish subordinate performance. Our theoretical predictions continue to hold if we redefine accuracy such that supervisors with higher $\left(1 - \frac{\eta}{2}\right)^2 \kappa$ are classified as accurate, and those with lower values as non-accurate. In contrast, when some supervisors are characterized by $\eta > 2$, those supervisors are so enormously incentivized to hoard talent that the sign of r^*_{TH} flips relative to r^*. Our predictions are likewise reversed when such supervisors are prevalent, or when η is sufficiently large for even a small subset of supervisors with $\eta > 2$, whereas our empirical results indicate otherwise.

Appendix C Optimal Threshold

In Section 2 we treat the promotion threshold as exogenous. This appendix outlines how the employer selects that threshold. Assume the employer observes the supervisor's reported evaluation r^* and then decides whether to promote the rated employee. Suppose the employer's payoff from promoting an employee is given by $q-c$, where $c > 0$. That is, promoting an employee with $q > c$ is profitable while promoting one with $q < c$ is costly. The payoff reflects the value of promoting high performers to more productive roles, and the potential incentive effects generated by promotion opportunities.

Suppose that there is only one type of supervisors. The employer knows the distributions and the supervisor's objective function, so she anticipates that $r^* = \frac{\sigma_s^2}{\sigma_q^2+\sigma_s^2}\overline{q} + \frac{\sigma_q^2}{\sigma_q^2+\sigma_s^2}\widehat{q}$. Since there is only one type of supervisors, the employer can perfectly recover the supervisor's local signal $\widehat{q}$ by computing $(r^* - \frac{\sigma_s^2}{\sigma_q^2+\sigma_s^2}\overline{q}) \times \frac{\sigma_q^2+\sigma_s^2}{\sigma_q^2} = \widehat{q}$. Then, the employer's expected profit by promoting the employee is

$$\mathsf{E}[\, q - c \mid \widehat{q}\,].$$

So, she promotes the employee if $\mathsf{E}[\, q \mid \widehat{q}\,] = \frac{\sigma_s^2}{\sigma_q^2+\sigma_s^2}\overline{q} + \frac{\sigma_q^2}{\sigma_q^2+\sigma_s^2}\widehat{q} > c$, which is after all equivalent to

$$r^* > c.$$

The optimal threshold is given as $t^* = c$. Notice that the threshold is independent of σ_s^2, so the optimal threshold is also c when there are more than one type of supervisors. This symmetry in the optimal thresholds are driven by the assumption that employer's costs of type-I and type-II errors are symmetric.

Appendix D Proofs for the Theoretical Predictions

Proof of Prediction 1

Recall from Equation (2) that the optimal evaluation is given by

$$r_s^* = \frac{\sigma_s^2}{\sigma_q^2 + \sigma_s^2}\overline{q} + \frac{\sigma_q^2}{\sigma_q^2 + \sigma_s^2}\widehat{q}.$$

Since the first term and the second coefficient are constants, we have

$$\mathsf{Var}[\, r_s^* \,] = \left(\frac{\sigma_q^2}{\sigma_q^2 + \sigma_s^2}\right)^2 \mathsf{Var}\,[\, q + \varepsilon_s \,] = \left(\frac{\sigma_q^2}{\sigma_q^2 + \sigma_s^2}\right)^2 \left(\sigma_q^2 + \sigma_s^2\right) = \frac{\sigma_q^4}{\sigma_q^2 + \sigma_s^2}.$$

The second equality holds because q and ε_s are independent. It is easy to see that if $\sigma_A^2 < \sigma_N^2$, then

$$\frac{\sigma_q^4}{\sigma_q^2 + \sigma_A^2} > \frac{\sigma_q^4}{\sigma_q^2 + \sigma_N^2}.$$

■

Proof of Prediction 2

Observe that in Equation (3), the argument inside $\Phi(\cdot)$ is increasing in σ_s^2. Since Φ is an increasing function, $\mathsf{Pr}(r_s^* \geq t)$ is decreasing in σ_s^2, which concludes the proof.

Proof of Prediction 3

The conditional expectation of $\widehat{q}$ given q is

$$\mathsf{E}[\, \widehat{q} \mid q \,] = q.$$

Substituting this into the definition of r_s^*, we get the conditional expectation of r_s^* given q:

$$\mathsf{E}[\, r_s^* \mid q \,] = \frac{\sigma_s^2}{\sigma_q^2 + \sigma_s^2}\overline{q} + \frac{\sigma_q^2}{\sigma_q^2 + \sigma_s^2}q.$$

Now consider the difference in conditional expectations for two performance levels q and q' such that $q' > q$, which is

$$\mathsf{E}[\, r_s^* \mid q' \,] - \mathsf{E}[\, r_s^* \mid q \,] = \frac{\sigma_q^2}{\sigma_q^2 + \sigma_s^2}(q' - q).$$

Since we assume that $\sigma_A^2 < \sigma_N^2$, we know that

$$\frac{\sigma_q^2}{\sigma_q^2 + \sigma_A^2} > \frac{\sigma_q^2}{\sigma_q^2 + \sigma_N^2}.$$

It follows that

$$\mathsf{E}[\, r_A^* \mid q' \,] - \mathsf{E}[\, r_A^* \mid q \,] > \mathsf{E}[\, r_N^* \mid q' \,] - \mathsf{E}[\, r_N^* \mid q \,],$$

which is equivalent to

$$\mathsf{E}[\, r_A^* \mid q' \,] - \mathsf{E}[\, r_N^* \mid q' \,] > \mathsf{E}[\, r_A^* \mid q \,] - \mathsf{E}[\, r_N^* \mid q \,].$$

■

Proof of Prediction 4

Let ρ be the correlation coefficient between q and r_s^*, and let ϕ denote the density function of the standard normal. By Theorem 22.5 of Greene (2003), we have

$$\mathsf{E}[\, q \mid r_s^* \geq t \,] = \overline{q} + \rho\, \sigma_q \cdot \frac{\phi\left(\frac{(t-\overline{q})\sqrt{\sigma_q^2+\sigma_s^2}}{\sigma_q^2} \right)}{1 - \Phi\left(\frac{(t-\overline{q})\sqrt{\sigma_q^2+\sigma_s^2}}{\sigma_q^2} \right)}.$$

Notice that

$$\begin{aligned}
\mathsf{Cov}[\, q, r_s^* \,] &= \mathsf{E}[\, (q-\overline{q})(r_s^* - \overline{q}) \,] \\
&= \mathsf{E}\left[(q-\overline{q}) \left(\frac{\sigma_s^2}{\sigma_q^2+\sigma_s^2}\overline{q} + \frac{\sigma_q^2}{\sigma_q^2+\sigma_s^2}\widehat{q} - \overline{q} \right) \right] \\
&= \mathsf{E}\left[(q-\overline{q}) \left(\frac{\sigma_q^2}{\sigma_q^2+\sigma_s^2}\widehat{q} - \frac{\sigma_q^2}{\sigma_q^2+\sigma_s^2}\overline{q} \right) \right] \\
&= \mathsf{E}\left[(q-\overline{q}) \left(\frac{\sigma_q^2}{\sigma_q^2+\sigma_s^2}(q+\varepsilon_s) - \frac{\sigma_q^2}{\sigma_q^2+\sigma_s^2}\overline{q} \right) \right] \\
&= \mathsf{E}\left[(q-\overline{q}) \left(\frac{\sigma_q^2}{\sigma_q^2+\sigma_s^2}(q-\overline{q}) + \frac{\sigma_q^2}{\sigma_q^2+\sigma_s^2}\varepsilon_s \right) \right] \\
&= \frac{\sigma_q^2}{\sigma_q^2+\sigma_s^2}\mathsf{E}\left[\, (q-\overline{q})^2 \,\right] + \frac{\sigma_q^2}{\sigma_q^2+\sigma_s^2}\mathsf{E}\left[\, (q-\overline{q})\, \varepsilon_s \,\right] \\
&= \frac{\sigma_q^4}{\sigma_q^2+\sigma_s^2}.
\end{aligned}$$

The last equality follows from the assumption that q and ε_s are independent. So, we have

$$\rho = \frac{\mathsf{Cov}[\, q, r_s^* \,]}{\sqrt{\mathsf{Var}[\, q \,]\mathsf{Var}[\, r_s^* \,]}} = \frac{\frac{\sigma_q^4}{\sigma_q^2+\sigma_s^2}}{\sigma_q \cdot \sqrt{\frac{\sigma_q^4}{\sigma_q^2+\sigma_s^2}}} = \frac{\sigma_q}{\sqrt{\sigma_q^2+\sigma_s^2}}.$$

Substitute this into the first equation and obtain

$$\mathsf{E}[\, q \mid r_s^* \geq t \,] = \overline{q} + \frac{\sigma_q^2}{\sqrt{\sigma_q^2+\sigma_s^2}} \cdot \frac{\phi\left(\frac{(t-\overline{q})\sqrt{\sigma_q^2+\sigma_s^2}}{\sigma_q^2}\right)}{1-\Phi\left(\frac{(t-\overline{q})\sqrt{\sigma_q^2+\sigma_s^2}}{\sigma_q^2}\right)}.$$

Clearly, the first fraction $\frac{\sigma_q}{\sqrt{\sigma_q^2+\sigma_s^2}}$ is decreasing in σ_s^2. Recall that we are assuming $t \geq \overline{q}$, so $\frac{(t-\overline{q})\sqrt{\sigma_q^2+\sigma_s^2}}{\sigma_q^2}$ is positive and increasing in σ_s^2. Notice that the last fraction $\frac{\phi(\cdot)}{1-\Phi(\cdot)}$ is the inverse Mills ratio, which is known to be strictly decreasing. The product of two positive-valued decreasing functions is also decreasing, so $\mathsf{E}[\, q \mid r_s^* \geq t \,]$ is decreasing in σ_s^2. Hence, we have $\mathsf{E}[\, q \mid r_A^* \geq t \,] > \mathsf{E}[\, q \mid r_N^* \geq t \,]$. ■

Appendix E Summary Statistics

Table E1: Summary Statistics by Sample

	Event-Study Sample			Full Sample		
Variable	Obs.	Mean	SD	Obs.	Mean	SD
Promotion	6512	0.14	0.34	17147	0.13	0.34
Numeric Grade	8481	2.92	1.47	19425	3.31	1.43
Age	8481	38.76	11.85	19425	40.93	10.88
Tenure	8481	14.85	12.52	19425	17.12	11.46
Standardized Evaluation Score	7889	0.03	0.93	18374	0.02	0.94
Education: Junior High School	8481	0.01	0.09	19425	0.01	0.07
Education: High School	8481	0.28	0.45	19425	0.24	0.43
Education: Community Collage	8481	0.02	0.13	19425	0.02	0.13
Education: Vocational School	8481	0.06	0.23	19425	0.06	0.24
Education: Collage of Technology	8481	0.09	0.29	19425	0.09	0.28
Education: University	8481	0.33	0.47	19425	0.35	0.48
Education: Master	8481	0.21	0.41	19425	0.23	0.42
Education: Doctor	8481	0.00	0.06	19425	0.00	0.07
Education: Others	8481	0.00	0.06	19425	0.00	0.04
Drinking Habit	6281	0.38	0.49	15474	0.40	0.49
Smoking Habit	6281	0.27	0.44	15474	0.27	0.44
Exercise Habit	6281	0.27	0.44	15474	0.27	0.44
Younger Cohort Dummy	8481	0.49	0.50	19425	0.49	0.50

Notes: Our data contains 19,425 observations of 1,977 unique employees. The full sample of our data contains data for the employees defined in section 3 for the period 2006-2019. We restrict the sample to conduct the event study in Section 4. We focus on a first supervisor switch during the observation periods and drop data when a second supervisor switch occurred.

Appendix F Full Sample Estimates

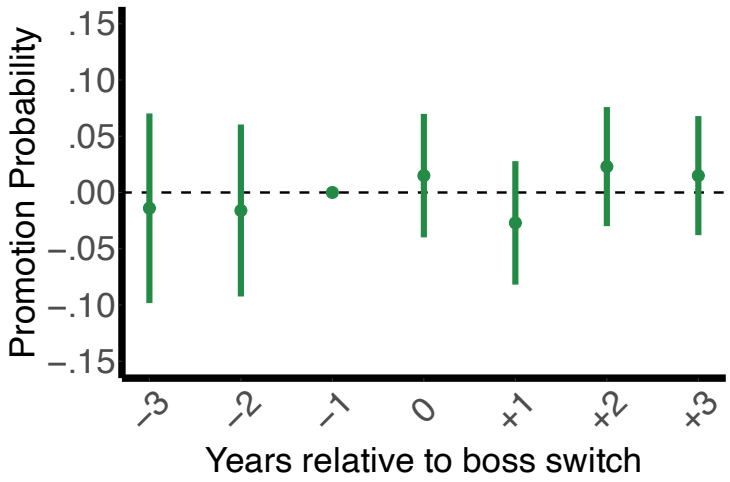


Figure F1: The Dynamic Effects of Accurate Raters on Employee Promotions (Full Sample)

Notes: See Section 4.2 for details about the model specification. The figure presents the estimation results from the event-study model. Each dot represents the estimated effect of reporting to an accurate-rating supervisor in each event period (estimate of β_k^D) using the full sample of male employees. The vertical segments represent the 95 percent confidence intervals. The standard errors are clustered at the employee and supervisor levels. Period 0 is the exact year when employees start reporting to new supervisors. Period -1 is omitted as the baseline in estimation, so the coefficient for period -1 is zero by construction.

Here, we present the estimates using the full sample of male employees. The econometric methods and model specifications are the same in Sections 4.2.1 and 4.2.2. Figure F1 shows the dynamic effects of switching to an accurate-rating supervisor on promotion probability. Unlike the restricted sample results, the estimated effects here are positive but smaller and not statistically significant in most periods. Figure F1 reports the estimates of Equation (8) using the full sample. Again, the estimates are positive but smaller. These results suggests that the positive relationship between supervisor accuracy and promotion outcomes is less precisely estimated when we do not condition on the timing and structure of supervisor changes.

There are several possible explanations for this pattern. First, the restricted sample focuses on employees observed within three years of their first supervisor switch and who continue working under the new supervisor, allowing for a cleaner identification of exposure to the focal supervisors. In contrast, the full sample includes employees with varying durations of supervisor relationships and potentially multiple supervisor switches, which introduces greater heterogeneity and noise.

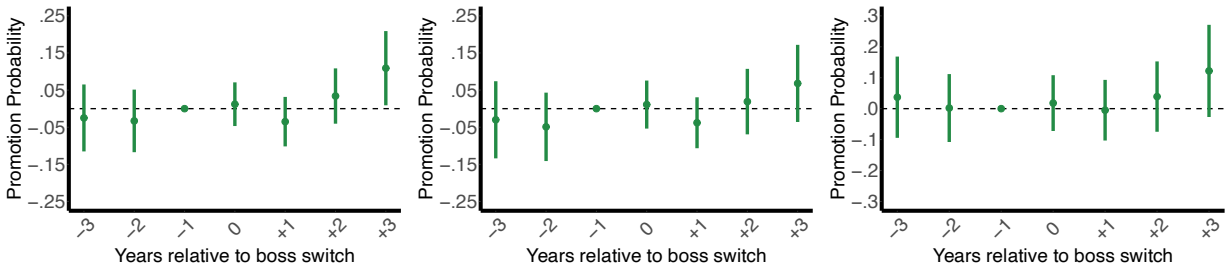
Table F1: Supervisor Evaluation Accuracy and Employee Promotions (Full Sample)

Outcome: Promotion	(1)	(2)	(3)	(4)
Supervisor Accuracy ($\widehat{\sigma}_{s(i,t)}$)	0.0323**	0.0163	0.0270**	0.0140
	(0.0127)	(0.0138)	(0.0127)	(0.0136)
Employee FEs	No	Yes	No	Yes
Year FEs	No	No	Yes	Yes
Age^2 & Tenure^2	Yes	Yes	Yes	Yes
Observations	17,147	17,147	17,147	17,147

Notes: See Section 4.2.2 for details on the model specification. The table shows the estimation results of regressions of promotions on supervisor accuracy using the full sample of male employees. The outcome is a binary indicator for whether the employee is promoted in a given year. The key independent variable is the square root of the estimated variance of supervisor ratings. Columns vary in included fixed effects. The numbers in parentheses are standard errors clustered at the employee and supervisor levels. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Second, even when employees are assigned to accurate supervisors, it may take time for the supervisor to learn about their subordinates' abilities and performance. Accurate evaluations are likely based on accumulated observations and interactions, which require time to form. Furthermore, even well-informed evaluations may take additional time to be reflected in formal promotion decisions, especially in organizations with structured or infrequent promotion cycles. These factors may weaken or delay the observable effects in the full sample, leading to the statistically insignificant estimates shown in the event-study graph.

Appendix G Robustness Check: Event-Study



(a) Leave-One-Out (b) Staggered DID (c) $N2A - N2N$

Figure G1: The Effect of Accurate-rating Supervisors with Certain Specifications

Notes: See Section 4.2.2 for details on the model specification. The figure presents the estimation results from the event-study model. Each dot represents the estimated effect of reporting to a accurate-rating supervisor in each event period. The vertical segments represent the 95 percent confidence intervals. The standard errors are clustered at the employee and supervisor levels. Period 0 is the exact year when employees start reporting to new supervisors. Period -1 is omitted as the baseline in estimation, so the coefficient for period -1 is zero by construction.

We check the robustness of our primary event-study result in three ways. First, one caveat of our proxy for supervisor accuracy (Equation (6)) is that, when we use the supervisor's accuracy measure as an independent variable in regressions, the residualized assessment score of the focal employee appears on both the left-hand side and in the construction of this right-hand side quantity. This mechanical correlation may bias inference, particularly in specifications using the full sample. We note that, however, our primary results that follow remain qualitatively the same when we use the following leave-one-out specification:

$$\widehat{\mathsf{Var}\left[r_s^*\right]}_{(j)} = \frac{1}{\sum_{t=1}^{T} |S_{s,t} \setminus \{j\}|} \sum_{t=1}^{T} \sum_{i \in S_{s,t} \setminus \{j\}} \left(\widehat{\eta}_{i,t}\right)^2,$$

where the variance is defined with respect to each observed pair of a supervisor and an employee. Panel (a) shows the estimation result, which is similar to the original result.

Second, we address the potential staggered nature of supervisor switches. We perform the estimation procedure developed by Sun and Abraham (2021). Panel (b) shows the estimation result. Although the estimate on period 3 is no longer statistically significant, the patterns of the point estimates are similar.

Another caveat is that our main specification does not consider the type of supervisor prior to a supervisor switch. Hence, our estimates confuse the effects of subordinates who changed from an accurate-rating supervisor to an accurate-rating supervisor ($A2A$) and those who changed from a nonaccurate-rating supervisor to an accurate-rating supervisor ($N2A$). Based on Cullen and Perez-Truglia (2023) and Minni (2023), we estimate the following regression equation:

$$y_{it} = \sum_{j \in J} \sum_{k \in \mathcal{K}} \beta_k^j D_{i,t-k}^j + \alpha_i + \tau_t + \varepsilon_{it}, \tag{G1}$$

where $j \in \{N2N, N2A, A2N, A2A\}$. Other expressions are the same as Section 4.2. We estimate the difference between β_s^{N2A} and β_s^{N2N}, which indicates the effect of reporting accurate-rating supervisor as specified in Equation (G1). Panel (c) shows the estimation result. Although the estimate on period 3 is no longer statistically significant, the patterns of the point estimates are similar.

These results are similar to those in section 4.2.1, albeit less precise, indicate that the implications remain unchanged based on our event study specification (Equation (7)).

Appendix H The Definition of Drinking Status

Table H1: The Distribution of Drinking Habits

	Amount			
Frequency	(a) 0–19 g	(b) 20–39 g	(c) 40–59 g	(d) 60+ g
(i) Hardly	Not (15%)	Not (1%)	Not (0%)	Drinking (0%)
(ii) Sometimes	Not (21%)	Not (19%)	Drinking (9%)	Drinking (4%)
(iii) Every day	Not (7%)	Drinking (14%)	Drinking (7%)	Drinking (3%)

Notes: This table specifies the definition of drinking workers. Workers are required to answer two questions about their drinking habits as part of the mandatory annual health check. One question asks the frequency of alcohol consumption. The possible answers are (i) hardly (or cannot drink), (ii) sometimes, and (iii) every day. The other question asks the average quantity of alcohol consumption on a day when they drink. The possible answers are (a) less than 20 grams, (b) 20 or more and less than 40 grams, (c) 40 or more and less than 60 grams, and (d) 60 grams or more in pure alcohol. The table defines drinking workers by mapping from two responses to either "Not" (non-drinking employees/bosses) or "Drinking" (drinking employees/bosses). The frequency of employee-year and boss-year observations for each cell are presented in brackets. The frequency is relative to the sum of the numbers of employee-year observations and boss-year observations.

To determine the drinking status of employees and their bosses, we use data from the annual health checks from 2015 to 2019. The health check records contain workers' responses to two questions about their drinking habits. One question asks about the frequency of alcohol consumption. The possible answers for this item are (i) hardly (or cannot drink), (ii) sometimes, and (iii) every day. The other item asks about the average amount of alcohol consumed on a day when they drink. The possible answers are (a) less than 20 grams, (b) 20 or more and less than 40 grams, (c) 40 or more and less than 60 grams, and (d) 60 grams or more in pure alcohol.[18]

We define drinking employees and bosses by the combination of their responses to the two drinking-related questions as specified in Table H1. We use the oldest drinking status records we have for each employee and each boss to classify drinking and non-drinking employees and bosses.[19] Since health check results are confidential and only available to a limited number of industrial health staffers, it is unlikely that workers have any incentives to report untruthfully for career concerns.

[18]The actual wording of the question uses a Japanese unit used to measure sake, *gō*. 1 *gō* amounts to approximately 180 milliliters of sake, which typically contains approximately 20 grams of pure alcohol. For reference, note that a glass of wine typically contain approximately 10 grams of pure alcohol.

[19]The estimation results are qualitatively unchanged when we use the latest records.

Appendix I Employee Engagement Survey

Responses are recorded using Likert scales ranging from three to six points, depending on the item and survey year. Values are coded so that higher values indicate more favorable responses. For analysis, all scores are normalized within each year. The questionnaire items are listed below. For ease of presentation, we assign a short label to each item, shown in brackets.

About Your Current Job

1. How do you perceive the amount of work you are currently handling? (Workload perception)
2. Compared to your role grade, do you feel that you are assigned a role appropriate to that grade? (Role-grade match)
3. Compared to your role grade, how do you feel about the current role and work assigned to you? (Difficulty of work)
4. Do you feel that your current job allows you to fully utilize your skills, knowledge, and abilities? (Skill utilization)
5. Are you given the authority and discretion necessary to carry out your job? (Authority delegation)
6. Are you interested in and motivated by your current job? (Interest & motivation)
7. Overall, how satisfied are you with your current job in terms of achievement and fulfillment? (Job satisfaction)

About Your Current Workplace

8. Is there a workplace atmosphere where people can freely express their opinions? (Open communication)
9. Is important work-related information shared among all members in your workplace? (Information sharing)
10. Does your workplace foster collaboration and teamwork to get work done?
11. Are tasks shared appropriately so that work does not fall disproportionately on individuals? (Team collaboration)
12. Is work arranged so that it can be completed within working hours? (Time efficiency)

13. Do you receive appropriate instructions from your supervisor based on sound judgment? (Supervisory guidance)

14. Does your supervisor provide development opportunities suited to your abilities and individuality? (Growth opportunities)

15. Were you satisfied with the year-end performance review with your supervisor last fiscal year (based on the role execution sheet)? (Review satisfaction)

16. How was the implementation of the beginning-of-year performance review with your supervisor this fiscal year (based on the role execution sheet)? (Beginning-of-year meeting)

17. Overall, how satisfied are you with your current workplace? (Workplace satisfaction)

About Your Current Company

18. Have you received explanations of the company-wide vision and strategy, and do you understand the content? (Understand the company's vision)

19. In your department, do you think there is adequate cross-departmental coordination necessary for work? (Inter-department collaboration)

20. Do you think the the company group will continue to demonstrate its unique strengths into the future? (View on the company's future)

Summary

21. Overall, how satisfied are you with working at the company group? (Company satisfaction)