

The Employer Bias: A Lab-based Study on Evaluating Employees' Self-Assessments*

Christine Alamaa¹ 

¹Swedish Institute for Social Research (SOFI), Stockholm University

November 22, 2025

[Latest version](#)

Abstract

In many organisational contexts, managers rely on limited information—such as employee self-evaluations or observable relative performance—when making decisions about task allocation, promotions, and compensation. While there is ample evidence on the role of individuals' often-biased self-representations, much less is known about how such expressions are judged by others. This study examines how managers interpret signals of employee performance and self-confidence, focusing on whether these interpretations differ systematically by manager gender and whether certain self-assessment biases are perceived as typically male or female. To address these questions, I conduct a laboratory experiment with 192 employers evaluating 576 employees. The results reveal substantial gender differences: female employers consistently perceive employees—particularly high performers—as about 0.2–0.3 standard deviations less confident than male employers. Providing explicit information on self-confidence inaccuracies strongly activates gender stereotypes: principals predominantly associate underconfidence with female employees and overconfidence with male employees. These findings underscore the presence of social biases in interpreting performance and self-confidence, with potential implications for gender dynamics and managerial decisions in organisations that rely on employee self-evaluations.

Keywords: Perception bias, gender stereotypes, employers, hiring and promotion decisions, lab experiment

JEL Codes: J01, J16, J50

*I am grateful for valuable comments from Erik Lindqvist, Jenny Säve-Söderbergh, Johannes Hagen, Max Brès, Louis-Pierre Lepage, Giulia Vattoune, and participants at the SOFI Brown Bag Series. I am also truly thankful to Tommaso Batistoni for invaluable coding assistance. Operative support in conducting the experiment was provided by the BLESS Lab at the University of Bologna. Financial support from the Institute for Evaluation of Labour Market and Education Policy (IFAU) is gratefully acknowledged.

Corresponding author: Christine Alamaa, christine.alamaa@sofi.su.se, Swedish Institute for Social Research.

1 Introduction

In professional environments, individuals’ self-assessments can influence their decisions to pursue tasks, seek promotions, and negotiate salaries. Self-confidence can motivate individuals toward ambitious professional paths. However, biases in self-assessments and representations can adversely affect career progress. While many scholars document systematic gender differences in self-promotions, often with men exhibiting greater overconfidence than women (Bengtsson et al., 2005; Lénárd et al., 2024), others emphasise their high task- and context-dependence (Bandiera et al., 2022; Clark and Friesen, 2009; Moore and Cain, 2007). Despite the established significance of self-confidence biases, far less research has examined how they are perceived by others—particularly by employers and managers (with Thoma, 2016, and a few others as exceptions).

Managers frequently rely on employee self-assessments, particularly when objective performance measures or direct observations are limited. Misreadings of employee qualities can lead to organisational inefficiencies through suboptimal task allocation and missed targets. Overly confident employees may underperform in tasks beyond their capabilities or receive compensation exceeding their productivity (see e.g., Malmendier and Tate, 2008; Barber and Odean, 2001). Conversely, underconfident employees risk being assigned tasks below their actual capacity (Babcock et al., 2017; Exley and Kessler, 2022), resulting in lost opportunities for both individuals and organisations. While recent research emphasises how gender stereotypes influence managerial decisions in strategic contexts such as hiring or promotion (Bohnet et al., 2016; Reuben et al., 2014), less attention has been devoted to understanding managerial perceptions of employee self-confidence signals in non-strategic contexts. In these settings, managerial perceptions of self-confidence may be influenced by social stereotypes—including gendered assumptions—rather than strategic motivations.

In this paper, I experimentally address this research gap by exploring two dimensions of managerial perceptions of employees’ relative performance and self-confidence. First, I examine how employers interpret observed employee performance ranks as indicators of employee self-confidence, and whether these interpretations vary by manager gender. Second, I investigate whether managers associate specific self-confidence inaccuracies—underconfidence and overconfidence—with employee gender. Managerial perceptions of employee self-confidence are difficult to analyse in the field as self-confidence typically interacts with other employee characteristics.¹ The anonymous laboratory environment allows me to isolate managers’ perceptual biases from both strategic behaviour and other confounding traits by varying only the salience of employee self-confidence.

I conduct a controlled laboratory experiment with 192 participants in the role of managers (principals), each evaluating three randomly matched employees, for a total of 576 assessments. In the experimental environment, employees complete a real-effort task to establish an objective but unrevealed performance rank. Then, employees assess their relative performance position. Meanwhile, principals first receive information on their employees’ objective performance ranks and complete an estimation task predicting the employees’ rank beliefs.² Thereafter, I introduce the principals’ information treatment, which

¹Examples of such traits include ambition, work ethic, likeability, and cooperativeness.

²More specifically, principals are asked to guess how many ranks the employee selected, which rank they considered most likely, and how much weight they assigned to it, among other aspects of the employee’s self-assessment. A full list of

comprises two conditions. In the *NoReveal* condition, principals receive no further information beyond employees' objective ranks. In the *Reveal* condition, principals also observe the rank in which employees expressed their strongest belief, allowing them to infer self-confidence. Afterwards, principals guess their employees' gender. By design, employees know what information principals receive, while to prevent gendered self-stereotyping, they remain unaware that it is used by principals for predictions and gender inference. This setup isolates the test of whether managers' gender stereotypes emerge in response to signals of employee self-confidence accuracy, beyond observing their objective performance.

The analysis proceeds in two steps. First, I compare how male and female principals interpret employees' ranks as signals of self-confidence, using questions asked before the experimental variation.³ Second, I analyse how principals infer employees' gender after observing their self-assessment accuracy (*Reveal*). This allows me to test whether signals of self-confidence type and accuracy trigger gender stereotyping in principals' evaluations of employees.

The first part of the analysis reveals systematic gender differences in managerial perceptions. Across both female and male principals, about 70% of employees are believed to rank themselves ahead of their actual performance (i.e. overconfident), about 20% to underplace their relative standing (i.e. underconfident), and about 10% to state their position accurately. These averages, however, uncover an underlying gender gap in how principals interpret rank signals. Female principals perceive employees as less confident than their male counterparts do. This tendency stems from female principals both under-attributing overconfidence and over-attributing underconfidence. Male principals, on the other hand, are more prone to believe that employees are accurate in their self-assessments. While the managerial gender gap in principals' inference of employees' self-confidence is moderate in size (about 0.2 standard deviations), it may be practically relevant. Over time, even small differences in perception may accumulate in organisational decision-making.

The second part of the analysis reveals managerial gender stereotyping. In the *Reveal* condition, principals observe employees' self-assessed rank alongside their actual rank. Then, principals observing an underconfident employee are about 18 percentage points more likely to guess this employee is female, whereas principals observing an overconfident self-assessment are about 11 percentage points less likely to do so. In other words, underconfidence is stereotyped as female and overconfidence as male. This association is strongest for top performers. By contrast, salient accurate self-assessments show no systematic association with employee gender. In sum, women are perceived to perform as well as men, but to have lower self-confidence. Such gender stereotyping appears conditional on the salience of self-confidence biases (*Reveal*), but not activated in the *NoReveal*, where only rank is observed.

Evidence directly on how managers interpret employees' confidence signals is scarce, although related research has only recently begun to examine how such signals of confidence and ability are judged in other contexts. My research extends the foundational insights of [Thoma \(2016\)](#), who show that underconfidence is systematically rewarded more positively than overconfidence, by examining how

the questions is provided in [Subsection 2.1.1](#).

³While all principals at this stage receive the same information, the random principal–employee matching ensures that female and male principals are equally likely to be paired with different combinations of employee ranks. This randomisation enables incentive-compatible comparisons across principal gender.

such perceptions vary by gender. This echoes the broader pattern of asymmetric evaluation by gender documented by [Sarsons \(2017\)](#); [Sarsons et al. \(2021\)](#), but in a context of medical outcomes rather than comparative self-assessments. My results also connect to [Barron et al. \(2025\)](#), who distinguish explicit from implicit discrimination in hiring: similarly, stereotypes in my experiment emerge only when self-assessment accuracy is revealed, not when performance alone is visible. From a theoretical perspective, my findings are consistent with [Bordalo et al. \(2016, 2019\)](#), who model stereotypes as distortions in belief formation—here underconfidence becomes coded as “female” and overconfidence as “male.” This aligns with broader organisational research showing that stereotypes can bias reviewers’ judgments even unconsciously when intending to be objective ([Heilman, 2012](#)). Finally, related hiring experiments show that stereotypes distort decisions even when candidates’ ability is held constant ([Reuben et al., 2014](#); [Coffman et al., 2021](#)).

The documented gender gaps in principals’ perceptions of confidence may have important implications in settings where managers must rely on employee self-evaluations. To illustrate, suppose principals generally assume that employees overestimate their ability and therefore mentally *deflate* self-signals when making judgments or assigning tasks. In such a scenario, female managers risk more pronounced over-assignment of overconfident employees while male managers risk more pronounced under-assignment of accurate or underconfident employees. In addition, male managers appear too prone to view self-confidence as precise and accurate, which may lead them to place excessive demands or provide insufficient support for miscalibrated employees. This underscores how evaluation formats matter: structuring assessments jointly rather than separately can reduce reliance on stereotypes, consistent with experimental evidence by [Bohnet et al. \(2016\)](#).

A further implication arises from the finding that gender stereotypes influence how principals interpret signals of self-confidence. If generalised to workplace settings, gendered readings of performance could bias task allocation: employees perceived as overconfident (and thus stereotyped as male) may be channelled into high-visibility tasks or leadership roles, while those perceived as underconfident (and thus stereotyped as female) may be relegated to routine or less challenging work. These patterns suggest that stereotypes can shape managerial perceptions, even when stereotyping is unintentional. Over time, such subtle mechanisms could perpetuate unequal opportunities, hinder women’s career advancement, and reinforce broader gender inequality within organisations.

The remainder of the paper is organised as follows: [Section 2](#) describes the experimental design and procedures; [Section 3](#) presents the empirical strategy; [Section 4](#) outlines the results and discusses implications ; and [Section 5](#) concludes with reflections on limitations and suggestions for future research.

2 Experimental Design

The experiment consists of two main parts (Part I, containing Task 1 and 2; and Part II). The experimental instructions use a standard labour market framing, are fully computerised, and appear privately on participants’ screens.⁴ Experimental parts were introduced to subjects sequentially,

⁴For example, participant roles (explained below), agent and principal, are framed as “Employee” and “Principal”; agents’ state their rank beliefs by selecting “contracts” to determine a “wage”-rate for their performance in a real-effort

detailed one after the other (see Appendix D for the full instructions, translated to English).⁵

Part I: Relative performance ranks and eliciting beliefs

The first part of the experiment contain two main tasks, Task 1 and Task 2 detailed below. In Task 1, all participants complete a real-effort task involving *decoding problems*. For each problem, participants decode a 5-digit number into a corresponding 5-letter text string. An example of a decoding problem is provided in Figure 1. Participants solve as many problems as possible within 4 minutes. A submitted answer cannot be changed, with each correct solution yielding one point and incorrect answers carrying no penalty. At the end of Task 1, participants are informed about their total number of attempted problems and make an incentivised guess about how many answers were correct. They subsequently receive feedback on their absolute performance, but no information about their relative performance or the performance of others.

Figure 1: Real effort task: an example of a decoding task problem

Decoding Task

Time remaining: **3:14**

| | | | | | | | | | | |
|----------------|---|---|---|---|---|---|---|---|---|---|
| Letter: | g | p | s | l | t | a | v | x | f | z |
| Key: | 6 | 3 | 1 | 0 | 7 | 9 | 2 | 4 | 5 | 8 |

Problem to solve:

42793

Enter your answer:

Attempts so far: 0

NOTES: The figure shows an example of a decoding problem. The displayed problem “42793” decodes to the solution “xvtap” (answers are not case-sensitive). The interface displays the remaining time on-screen and increases the count of attempted problems after each submitted answer. No other feedback is provided.

The experiment employs two participant roles, type A and type B. In each experimental session of 24 participants, 6 participants are randomly assigned to the role of *principal* (type A), and 18 to the role of *employee* (type B). Three employees are randomly matched into a group with one principal, without any strategic wage or hiring decisions. The 18 employees are ranked from 1 to 18 according to their performance in Task 1. Rank 1 is assigned to the employee with the highest performance, rank 2 to the employee with only one other employee scoring better, and so forth, until rank 18 is assigned to the employee with the lowest performance (i.e with 17 others scoring better). This ranking rule implies that

task, *et cetera*.

⁵Before beginning each experimental part, participants completed a short, unpaid trial task (2 and 4 minutes respectively), followed by a comprehension test covering both Part I and Part II.

all employees with a score tie share the better rank (lower number).⁶In Task 2, employees undertake a self-assessment task designed to determine their earnings based on their absolute performance in Task 1 and to elicit their rank beliefs. Employees receive an endowment of 19 units of a non-transferable, non-storable experimental currency (ECU), which they must allocate fully across *one* or *more* of the possible ranks (from rank 1, the highest, to rank 18, the lowest).

Employees complete this allocation in two steps: first by selecting ranks, and second by deciding how many ECUs to assign to each selected rank. Employees may revise and restart their rank selections (step 1) during the second step. Employees must allocate all 19 ECUs with a unimodal distribution, meaning that one rank must be assigned at least one more ECU than any other. These constraints facilitate obtaining a clearly interpretable measure of the employees' beliefs about their most likely actual performance rank.

Experimental payoff and incentive structure. Employees' earnings depend directly on their accuracy in allocating ECUs to their true performance rank. Specifically, employees allocate their entire endowment of 19 ECUs across one or more ranks (from 1 to 18). If an employee allocates ECUs to their true rank, their earnings are calculated as the number of allocated ECUs multiplied by their task score. Allocating no ECUs to their actual rank results in zero earnings. Allowing employees to allocate ECUs across multiple ranks explicitly accommodates their monetary risk preferences, enabling them to express uncertainty about their true rank clearly and without distortion. Principals' earnings are directly based on their employees' earnings, by default equal to one-third of the earnings of each of their three employees. This pay scheme ensures that principals directly benefit from their employees' accurate self-assessments. This alignment minimises potential influences from other-regarding concerns (e.g., altruism, inequality aversion) or "shying-away" effects, which could arise if employees believed, for example, that only the best-performing employee determined their principal's earnings.⁷ At the conclusion of the experiment, ECUs are converted into Euros at a fixed rate of 1 Euro per 10 ECUs.⁸ The next subsection introduce Part II of the experiment separately for employees and principals.

2.1 Part II: Information Treatments and Principal Decisions

After concluding Part I, participants proceed to Part II, which introduces varying information conditions separately for employees and principals. Employees first receive feedback about their own performance outcomes. Subsequently, principals receive information about their three assigned employees. The experimental design originally contained four distinct employee-principal information conditions. For

⁶For example, if the two highest-scoring employees both have 20 correct answers, both receive rank 1. The next-best employee is then assigned rank 3, since two others scored better. If no other ties occur, the lowest-ranked employee will still receive rank 18.

⁷Additionally, at the end of principals' Part II (outlined in 2.1.1), they can privately choose whether to earn their payoff based on all three employees (the default) or entirely on one selected employee. Crucially, employees remain unaware of this principal choice. Since employees have no knowledge of principals' self-serving strategic choice, the rank-belief elicitation remains free from employee considerations related to anticipated managerial reactions. The analysis of this principal choice is beyond the scope of this paper.

⁸The experiment consisted of three identical experimental rounds, repeating Task 1 and 2 three times. In this paper I analyse yet only the first round of employer-employee decisions. One round was randomly selected for payment and announced to the participant at end of the experiment.

analytical clarity and simplicity, these four conditions are pooled into two main principal treatment groups, reflecting the perspective and informational differences relevant to principals.

Employees' Information Treatments: Feedback Conditions

After completing Tasks 1 and 2, employees receive one of two types of task performance feedback. All employees privately receive a summary feedback screen that repeats their task score and their selected ranks ordered by their allocated ECUs, with the rank receiving the largest allocation highlighted and denoted their “most preferred rank.” Furthermore,

- Employees receiving only this baseline summary feedback, with no information about their actual performance rank, constitute the feedback control conditions, labelled: *NoFeedback* and *OnlyReveal*.
- Employees receiving the baseline summary feedback plus explicit information about their actual performance rank, constitute the feedback treatment conditions, labelled: *OnlyFeedback* and *FeedbackReveal*.

Subsequently, employees disseminate personal performance information to their respective principals *via* a submission form. Importantly, *within* the feedback control group, as well as the feedback treatment group, the information received by employees is identical. These conditions differ only in terms of the information subsequently provided to principals, as detailed in the next subsection.

Principals' Information Treatments: NoReveal and Reveal Conditions

For principals, Part II serves two purposes: first, to reveal information about their employees, and second, to complete an incentivised survey comprising six questions regarding employees' self-assessed performance ranks from Part I, Task 2. Initially, all principals receive baseline information consisting of their three employees' actual performance ranks. Principals then answer the first five survey questions, which remain unaffected by subsequent treatment variations. Following these questions, principals are randomly assigned into two distinct treatment conditions, which differ in whether principals receive additional information about their employees' self-assessments:

- **NoReveal (*NoFeedback*, *OnlyFeedback*):** Principals receive no information about their employees beyond their actual ranks (baseline). Thus, principals have no insights into employees' self-assessment accuracy regarding their most preferred rank.
- **Reveal (*OnlyReveal*, *FeedbackReveal*):**⁹ Principals receive, in addition to the baseline information, employees' stated preferred ranks, highlighting self-assessment deviations as either accurate, overestimations, or underestimations, and the magnitude of these deviations.

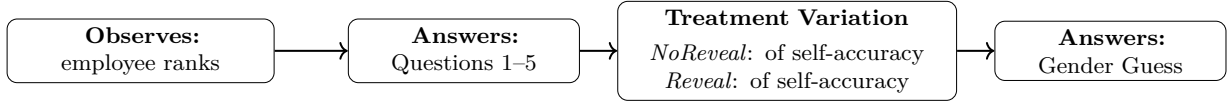
After receiving their respective treatment information, principals complete the sixth and final question of the incentivised survey, as detailed next (in [Subsection 2.1.1](#)). From the principals' perspective, employee-side feedback variations differ only in whether the employees receive private feedback about

⁹Importantly, principals in the *OnlyReveal* condition receive information about the true performance ranks of their three employees, even though the employees themselves remain unaware of their own ranks.

their true performance ranks and thus directly observe their own self-confidence accuracy.¹⁰ For analytical clarity and improved statistical power, I pool the four employee-principal treatment conditions (described above) into two distinct information conditions: **NoReveal** and **Reveal**. This pooling assumes that any differences in employees’ anticipation of private feedback are minor from the viewpoint of principals’ perceptions.¹¹ Further methodological details and identification concerns related to these treatment variations are discussed in [Section 3.3](#).

[Figure 2](#) summarises the timing of the four steps concerning the principals’ survey and their information treatment conditions, described next.

Figure 2: Principals’ decision module timeline



2.1.1 Principals’ incentivised survey

After receiving the objective performance rank of their three assigned employees, principals complete an incentivised survey, consisting of six questions.¹² Each question explicitly indicated whether principals should provide answers separately for each employee or for only one employee. The survey interface included built-in constraints to guide and remind principals about the logic underlying employees’ rank assessments. For instance, if a principal indicated that an employee allocated 10 ECUs to their most preferred rank, the interface constrained the allocation to a maximum of 9 ECUs for any other ranks.

- The questions principals answer are as follows:

1. **Number of ranks chosen by employee:** Principals guessed how many of the 18 possible ranks the employee allocated ECUs across (range: 1-18).

[10 ECUs per correct answer (1-3)].

“How many (of the 18 possible) contract(s) did the Employee allocate ECUs to?”

2. **Employee’s most preferred rank:** Principals guessed the employee’s modal (most preferred) rank (range: 1-18). [10 ECUs per correct answer (1-3).]

“Which was the Most Preferred Contract (MPC), i.e., the number of the contract?”

3. **ECUs allocated to the most preferred rank:** Principals guessed how many ECUs the employee allocated to their most preferred rank (conditional range, based on earlier guesses). [10 ECUs per correct answer (1-3)].

“How many ECUs was put on the Most Preferred Contract (MPC)?”

¹⁰By design, the employee-principal information conditions is *common knowledge* to both employees and principals, also before employees make their choices in Part I. Principals might therefore expect that employees, anticipating either a *feedback control* or a *feedback treatment* condition, could adjust their choices accordingly.

¹¹Balance tests confirming that principals do not systematically anticipate differences between these employee-side variations are reported and discussed further in [Section 3.3](#).

¹²The survey originally included two additional questions, not analysed in this paper, detailed in [Appendix A](#).

4. **ECUs allocated to the employee’s actual rank:** Principals guessed how many ECUs the employee allocated to the rank corresponding to their actual performance (conditional range, based on earlier guesses). [10 ECUs per correct answer (1-3)].

“How many ECUs were put on the contract with the same number as the actual rank?”

5. **Employee with highest earnings:** Principals guessed which employee (1, 2, or 3) achieved the highest earnings, defined as ECUs allocated to actual rank multiplied by task score. [Rewarded with 10 ECUs].

“Which of your Employees (1-3) do you think had the highest earnings (wage \times score)?”

- After these five questions, the treatment information (Reveal or NoReveal) was provided. Principals then answered one final question influenced by the treatment information:

6. **Gender guess (post-treatment):** Principals guessed the gender of each of their three employees (range: male/female). [10 ECUs for correct combination].

“Guess the gender of your employees by selecting Man/Women in the drop-down list below.”

This sequential ordering of these survey questions enables a descriptive analysis of gender differences among principals in their initial perceptions of their employees’ self-confidence (based on outcomes from employees’ rank self-assessments in Part I, Task 2). Importantly, only one question—the gender guess—is directly influenced by the treatment variation. This design thus allows for the causal analysis of how principals’ exposure to employee self-assessment information shapes their gender-based stereotyping of employees’ self-confidence signals.

2.2 Administration and experimental procedures

The data generation process was designed for a laboratory experiment: conducted in the Bologna Laboratory for Experiments in Social Science (BLESS) at the University of Bologna during 2023-2024. The experimental design is part of a larger experiment (analysing the behaviour of the employee). Yet, main hypotheses for this study were pre-registered at the Open Science OSF Preregistries.¹³ Participants were recruited through ORSEE (Greiner, 2015) admitting two different lists, one for men and one for women (conditioning on their registered gender).¹⁴ The experimental software was programmed in oTree[®] (Chen et al., 2016). Random assignments used for *e.g.* groups, treatments or random outcomes like lottery draws, were directly integrated in the software’s application code. The experimental instructions were fully computerised and provided in Italian. The entire experiment took about one hour to complete. The procedure began with reading the instructions, followed by a 2-minute decoding task trial. Participants then tried out the self-assessment task and completed a comprehension test, which took about 25 minutes. The larger experiment, consisting of three identical rounds (of Task 1 and 2 etc.), was completed in an additional 35 minutes. Principals’ average earnings (excluding a five or a ten euro show up fee) were 8.7 euros. One of the three experimental rounds was

¹³Alamaa, C. (2023, June 11), “To observe a bias with bias - An experiment on the signal value of other’s self-assessments” (previous working title) and can be retrieved from <https://osf.io/jgdqz> after June 29 2025.

¹⁴All subjects provide an informed consent *via* a Consent Form, before participation. The Consent Form apply with the EU GDPR Law (2016/679) regarding storage of Personal Data.

randomly selected for payment and announced in the end of the experiment. Earnings were denoted in experimental currency units that was converted to euro at rate 0.1 (*i.e.*, 10 ECUs correspond to 1 euro).

3 Data and Empirical Strategy

This section describes the dataset and empirical methods. [Section 3.1](#) details the sample and descriptive statistics, [Section 3.2](#) outlines the empirical approach and variables, and [Section 3.3](#) discusses identification and balance checks.

3.1 Sample and Descriptive Statistics

The analysis is based on experimental data collected from 40 laboratory sessions, involving 192 principals (96 males, 96 females) acting as employers, each matched one to three with 575 employees, and thus covering a total of 767 experimental subjects. Principals were randomly assigned to two treatment conditions: half (96 principals) to the *NoReveal* condition with 287 employees, and half (96 principals) to the *Reveal* condition with 288 employees.¹⁵ Each treatment condition maintains gender balance, with exactly 48 female and 48 male principals.

[Table 1](#) summarises the employee sample distribution by principal gender and treatment condition, reflecting the number of principal decisions recorded in each subgroup. The number of principals per gender-treatment subgroup is indicated in parentheses in the table header. The *NoReveal* condition includes 144 male and 143 female employees, while the *Reveal* condition includes 145 male and 143 female employees.¹⁶

Table 1: Sampled employees by principals' gender and treatment condition

| All (192) | | Male (96) | | Female (96) | |
|---------------|-------------|---------------|-------------|---------------|-------------|
| 575 | | 287 | | 288 | |
| NoReveal (96) | Reveal (96) | NoReveal (48) | Reveal (48) | NoReveal (48) | Reveal (48) |
| 287 | 288 | 144 | 145 | 143 | 143 |

NOTES: The table shows the final sample of employee observations (employee outcomes or principals guesses), categorised by principal gender and gender-treatment condition. The number of principals per subgroup is in parentheses.

To confirm the effectiveness of randomisation across treatment conditions, I perform covariate balance tests on principals' characteristics, presented in [Table B.1](#) in [Appendix B](#). A minor imbalance appears in principals' age (p -value = 0.003), driven by two outliers aged 40 or older. However, the remaining

¹⁵Within these conditions, principals were evenly distributed across four subgroups: *NoFeedback* (48 principals), *OnlyFeedback* (48 principals), *OnlyReveal* (48 principals), and *FeedbackReveal* (48 principals).

¹⁶The minor gender imbalance resulted from recruitment-related issues: one observation was excluded due to a mismatch between database registry and declared gender, and one experimental session oversampled a male participant due to challenges in recruiting sufficient female participants.

covariates are well balanced across treatments (p -values generally exceeding 0.05), ensuring robust causal inference.

Employees’ descriptive characteristics and beliefs regarding their ranks are summarised in Table B.2. Female employees systematically select worse preferred ranks than male employees ($p < 0.01$), reflecting significant gender differences in self-confidence measures as rank and rank belief precision. These gender differences remain consistent across treatment conditions, providing a stable foundation for subsequent analyses.

Table 2 summarises principal-level averages of three key characteristics describing the ranks of their assigned employees: the average employee rank per principal, the range of employee ranks per principal (highest minus lowest rank), and an indicator for any rank-ties among a principal’s three employees. The descriptive statistics are presented separately by principals’ treatment conditions (*Reveal* vs. *NoReveal*) and gender.

Table 2: Employee rank characteristics per principal, by treatment and principal gender

| | Scale | All | | | | NoReveal | | | | Reveal | | | | <i>ttest</i> |
|----------------------------------|-------|----------------|----------------|----------------|-----------------------|----------------|----------------|----------------|-----------------------|----------------|----------------|----------------|------------------------|---------------------------|
| | | All (1) | Male (2) | Female (3) | <i>p</i> -val. (4) | All (5) | Male (6) | Female (7) | <i>p</i> -val. (8) | All (9) | Male (10) | Female (11) | <i>p</i> -val. (12) | <i>p</i> -val. (5)-(9) |
| <i>Employees averages</i> | | | | | | | | | | | | | | |
| Rank avg. | 2-17 | 8.88 (2.77) | 8.60 (2.59) | 9.16 (2.93) | 0.161 | 9.01 (2.73) | 8.57 (2.60) | 9.45 (2.81) | 0.114 | 8.75 (2.83) | 8.63 (2.60) | 8.87 (3.06) | 0.676 | 0.511 |
| Rank range | 0-17 | 9.71 (3.97) | 9.71 (4.04) | 9.72 (3.93) | 0.986 | 9.82 (3.88) | 9.88 (4.04) | 9.77 (3.76) | 0.896 | 9.60 (4.08) | 9.54 (4.08) | 9.67 (4.13) | 0.882 | 0.704 |
| Rank tie | 0/1 | 0.22 | 0.26 | 0.19 | 0.226 | 0.22 | 0.25 | 0.19 | 0.459 | 0.23 | 0.27 | 0.19 | 0.331 | 0.863 |
| <i>N</i> | | 192 | 96 | 96 | 192 | 96 | 48 | 48 | 96 | 96 | 48 | 48 | 96 | 192 |

NOTES: The table presents principal-level averages of employee ranks, the range of ranks among a principal’s three employees, and an indicator for rank ties (among the top or bottom performers). Columns 1–4 report statistics for the full sample and separately by treatment conditions (*Reveal* in columns 5–8, *NoReveal* in columns 9–12) and principal gender. The final column reports p -values from t -tests for “Rank avg.” and “Rank range”, and a χ^2 -test for “Rank ties”, comparing the treatment conditions. Standard deviations are shown in parentheses.

Overall, principals manage employees with a mean average rank close to the mid-point of possible ranks (8.88). Substantial variation in employee ranks is evident, with an average rank range of 9.71 ranks per principal. Approximately 22 percent of principals have at least one tie in the ranks among their employees. Crucially, no significant differences across treatments or between male and female principals are observed for these characteristics, ensuring a balanced baseline for further analyses.

3.2 Empirical Strategy

The empirical strategy addresses two primary objectives. First, it explores observational gender differences in principals’ initial perceptions of employee self-confidence based on observed relative performance. Second, it examines the causal effect of introducing the treatment condition on principals’ gender stereotyping in interpreting employees’ self-confidence signals.

Gender Differences in Interpreting Relative Performance. For this first objective, the dependent variables (Y_{ij}) represent principals' evaluations regarding employees' stated rank beliefs. Specifically, principals' evaluations are measured by their guesses about the distribution of employees' rank selections, the most preferred rank selected by employees, and the precision of employees' confidence in their rank placements. Gender differences in principals' evaluations, for this set of outcomes, are assessed by comparing means using t -tests across male and female principals.¹⁷ A particular attention is paid to principals' perceptions of employee self-confidence, defined as the accuracy of an employee's self-assessment, comparing their true rank with their most strongly chosen rank ("most preferred rank"). The primary independent variables of interest include principal gender (*femPrincipal*), conditioning on controlling for employee performance rank (*Rank emp*). The model specification for this analysis is:

$$Y_{ij} = \alpha + \beta_1 \text{femPrincipal}_i + X_i' \gamma + \epsilon_{ij}, \quad (1)$$

where Y_{ij} represents principal i 's evaluation of employee j based solely on observed employee ranks. Estimations are performed using Ordinary Least Squares (OLS), with standard errors clustered at the principal level. Additionally, analyses separately examine employees that performed as the upper half ranks (ranks 1–9) and lower half ranks (ranks 10–18) to explore effects stemming from evaluating different performance level types. Control variables (X_i) include principals' background characteristics (age, education level, risk preferences) and employee attributes (actual performance ranks, preferred ranks), ensuring robustness and reducing omitted variable bias.

Interpreting self-confidence as gendered types. For the second objective, the dependent variable (Y_{ij}) captures principals' beliefs about the gender of employees, indicating the presence of gender stereotypes in interpreting employees' self-confidence types. This dimension specifically explores how observing employees' self-assessment biases (underconfidence, accurate confidence, overconfidence) under the *Reveal* condition influences gender-based stereotyping. The general form of the regression is:

$$\begin{aligned} Y_{ij} = & \alpha + \beta_1 \text{femPrincipal}_i \\ & + \beta_2 (\text{Reveal}_i \times \text{Underconf.}_i) + \beta_3 (\text{Reveal}_i \times \text{Accurate}_i) + \beta_4 (\text{Reveal}_i \times \text{Overconf.}_i) \\ & + X_i' \gamma + \epsilon_{ij}, \end{aligned} \quad (2)$$

where the primary independent variables test gender differences (*femPrincipal*), treatment effects and the interaction with the observed type of self-confidence bias (under-, accurate-, or overconfidence). Controls (X_i) include principal and employee characteristics to reduce potential omitted variable bias and enhance the robustness of causal interpretations.

The two objectives of the empirical analysis distinguishes between descriptive analyses, conducted on the initial five incentivised survey questions answered prior to treatment exposure, and causal analyses, focusing on the gender guess question answered after exposure to treatment information.

¹⁷ t -tests are used for discrete or continuous outcomes.

3.2.1 Variables Description

To perform the empirical analysis, I utilise variables derived from both employee tasks (1 and 2) and the principals' incentivised survey task (Part II).

The **dependent variables** capturing principals' evaluations of employees' rank beliefs include:

- **Rank belief distribution (Rank diversification), “uncertainty” [scale: 1–18]:** The total number of distinct ranks selected by an employee when allocating their belief weights in Task 2, indicating uncertainty about their true rank.
- **Modal rank belief (Preferred rank), “best guess” [scale: 1–18]:** The rank to which the employee allocates the largest number of ECUs, representing their most strongly held belief about their true rank.
- **Strength of best guess (Belief weight on preferred rank), “precision” [scale: 1–19]:** The number of ECUs an employee allocates to their preferred rank, capturing the strength or precision of their rank belief.
- **Strength of accurate guess (Belief weight on actual rank), “accuracy” [scale: 0–19]:** The number of ECUs an employee allocates to the rank corresponding exactly to their actual performance rank.
- **Self-placement accuracy (self-confidence), [scale: -17–17]:** The difference between the employee's actual performance rank and their most preferred rank, where negative values indicate underconfidence (the employee places themselves worse, with a higher numerical rank, than their true rank), and positive values indicate overconfidence (the employee places themselves better, with a lower numerical rank, than their true rank).
- **Gender guess (Female), “stereotyping” [binary scale: 0/1]:** Principals' guesses about the employee's gender, coded as 1 if guessed female and 0 if guessed male.

The **independent variables** of primary interest include the treatment conditions, principal gender, and employee self-confidence types:

- **Principal's gender (*femPrincipal*), “female principal” [binary scale: 0/1]:** Indicator variable coded 1 if the principal is female, and 0 otherwise.
- **Experimental treatment indicator (*Reveal*), [binary scale: 0/1]:** Indicator variable coded 1 if the principal observes the employee's preferred rank (Reveal condition), and 0 otherwise (NoReveal condition).
- **Self-confidence types (Underconfidence; Accurate confidence; Overconfidence), [binary scales]:** Categorical variables coded as indicators (1 if true, 0 otherwise) based on whether employees' preferred ranks are above (underconfidence), exactly equal to (accurate confidence), or below (overconfidence) their true ranks.

Two categories of **control variables** are included, covering both principal and employee characteristics:

Principal-level control variables:

- *Age*: Principal’s age in years.
- *Higher education*: The number of completed years of academic studies in higher education.
- *Risk-tolerance*: A binary indicator coded 1 if the participant selected the riskier option in a post-treatment lottery, and 0 otherwise.
- *Field of studies*: Categorical variables indicating principals’ academic background:

STEM major: Indicator variable coded 1 if the principal’s major is in science, technology, engineering, or mathematics, and 0 otherwise.

LLL major: Indicator variable coded 1 if the principal’s major is in languages, literature, or linguistics, and 0 otherwise.

Employee-level control variables:

- *Actual rank (employee)*: The employee’s true performance rank from the real-effort task.

3.3 Identification Strategy and Balance Checks

The experimental design enables causal identification of how revealing employee self-assessment accuracy influences principals’ gender stereotyping. Principals are randomly assigned to the *NoReveal* and *Reveal* conditions, ensuring internal validity provided randomisation successfully balances principals’ initial perceptions across these conditions.

One concern, particularly in the *Reveal* condition, is that principals might anticipate employees’ behaviour changing when their self-assessments become observable. If principals believe employees adjust their stated self-confidence due to social exposure, then their subsequent inferences about employee gender may reflect these anticipations rather than genuine gender stereotyping regarding self-confidence. Thus, a critical consideration is whether principals perceive female and male employees as systematically differing in their behavioural adjustments to anticipated social scrutiny.

Separate analyses confirm that employees indeed adjust their self-assessments in response to anticipated principal observation, indicating socially motivated behavioural changes. However, initial balance tests (Table B.3) suggest that principals do not systematically anticipate such employee adjustments differently across treatment conditions. The balance tests explicitly compare principals’ initial guesses about employee rank diversification, preferred ranks, and belief weights for both preferred and actual ranks—overall, and separately by principal gender. No statistically significant differences emerge in principals’ baseline guesses between the *NoReveal* and *Reveal* conditions. This support successful randomisation and ensure that subsequent gender stereotyping results genuinely reflect treatment-induced changes rather than pre-existing perceptual differences. To ensure the validity of pooling employee conditions into the two principal treatment groups (*NoReveal* and *Reveal*), I conducted balance tests (see Table B.4; B.5; B.6 in Appendix B). These compare principals’ baseline guesses between principal-side conditions (*NoReveal* vs. *Reveal*), separately within each employee treatment and by principal gender, generally reveal no systematic differences (pairwise *t*-tests, nearly all $p > 0.05$).

One exception appears in principals’ guesses of ECUs allocated to employees’ preferred ranks within the *NoFeedback* condition, showing a marginally significant difference ($p = 0.052$ overall), driven by female principals ($p = 0.039$). Although noteworthy, this single finding does not substantially undermine the validity of pooling conditions into the principal treatment groups *NoReveal* and *Reveal*. These comprehensive balance and robustness checks validate the randomisation and internal validity of the experimental design, ensuring reliable causal inference regarding how revealing employees’ self-assessment accuracy affects principals’ gender-based stereotyping.

4 Results and discussion

In this section, I present the key empirical findings derived from the experiment. First, I highlight gender differences in principals’ perceptions of employee behaviours and rank self-assessments (in [Section 4.1](#)). I then examine principals’ gender differences regarding their perceptions of employee self-confidence (outlined in [Section 4.2](#)). Finally, in [Section 4.3](#) I analyse principals’ gender stereotyping as revealed by their gender guesses.

4.1 Descriptive Gender Differences in Principals’ Perceptions

[Table 3](#) summarises principals’ guesses about employees’ rank beliefs by principal gender, and also displays the differences between principals’ guesses and employees’ actual choices (rows labelled *diff. (princ.-emp.)* and *diff. (emp.-princ.)*), respectively.

Female principals perceive employees as more uncertain about their ranks than male principals do. Specifically, female principals guess that employees report significantly more ranks as potentially correct (average 7.32 ranks) compared to male principals (average 5.65 ranks; $p < 0.001$). Compared to the actual behaviour of employees specifically assigned to them, female principals significantly overestimate employee uncertainty, guessing approximately two ranks more than their employees actually select (one-sample, two-sided t -test, $p < 0.001$). Male principals exhibit a much smaller discrepancy (approximately 0.35 ranks), which is not statistically significant (one-sample, two-sided t -test, $p = 0.12$). The difference between female and male principals’ accuracy in perceiving employee uncertainty is statistically significant (two-sample, two-sided t -test, $p < 0.001$).

Female principals predict that employees choose significantly worse ranks (higher numerical ranks; average 7.69) as their “best guess”, compared to male principals (average 6.52; $p = 0.001$). Compared to the actual preferred ranks of employees specifically assigned to them, female principals significantly overestimate employee choices, predicting ranks that are approximately 1.84 positions worse (one-sample, two-sided t -test, $p < 0.001$). Male principals exhibit minimal discrepancy (approximately 0.14 ranks), which is not statistically significant (one-sample, two-sided t -test, $p = 0.58$). The difference between female and male principals’ accuracy in predicting employees’ preferred ranks is highly statistically significant (two-sample, two-sided t -test, $p < 0.001$), indicating that female principals hold notably more pessimistic perceptions of employees’ rank preferences.

Principals’ guesses regarding employees’ belief precision (ECUs invested in their most preferred rank) and belief accuracy (ECUs invested in their actual rank) show no statistically significant gender

differences. Female and male principals similarly perceive the precision and accuracy of employees' expressed confidence. However, both female and male principals significantly overestimate the actual level of employees' belief precision and accuracy, with large and statistically significant discrepancies within each principal gender group (one-sample, two-sided t -tests, all $p < 0.001$). The results suggest notable gender differences among managers (principals) in their perceptions of employee behaviour, potentially influencing workplace decisions and task assignments. Female principals appear to perceive employees as more uncertain and pessimistic regarding their abilities and preferred tasks, often overestimating uncertainty and underestimating employees' self-assessed capabilities.

Table 3: Principals' guesses on employees' rank beliefs, by principal gender

| | | Pooled treatments (<i>NoReveal</i> and <i>Reveal</i>) | | | |
|--|--------------|---|----------------|----------------|-----------------------|
| | Scale | All (1) | Male (2) | Female (3) | <i>p</i> -val. (4) |
| <i>Principals' guesses on employees</i> | | | | | |
| Rank diversification | 1-18 | 6.49 (4.54) | 5.65 (3.75) | 7.32 (5.09) | <0.001 |
| <i>diff. (principal-all emp's.)</i> | -16-16 | 1.20 | 0.35 | 2.06 | <0.001 |
| Preferred rank | 1-18 | 7.11 (4.41) | 6.52 (4.15) | 7.69 (4.59) | 0.001 |
| <i>diff. (all emp's.-principal)</i> | -17-17 | -0.85 | 0.14 | -1.83 | <0.001 |
| Belief weight—preferred rank | 2-19 | 8.14 (4.05) | 8.24 (3.74) | 8.04 (4.35) | 0.550 |
| <i>diff. (principal-all emp's.)</i> | -16-16 | 1.17 | 1.15 | 1.18 | 0.943 |
| Belief weight—actual rank | 0-19 | 3.41 (3.29) | 3.68 (3.52) | 3.14 (3.02) | 0.052 |
| <i>diff. (principal-all emp's.)</i> | -19-19 | 1.86 | 2.07 | 1.64 | 0.118 |
| <i>Emp. earned the most</i> | | | | | |
| Correct guess (emp. top) | 0/1 | 0.35 | 0.37 | 0.33 | 0.650 |
| | <i>N=192</i> | 67 | 35 | 32 | |
| Highest rank | 0/1 | 0.65 | 0.57 | 0.72 | <0.001 |
| <i>Share correct</i> (highest) | 0/1 | 0.75 | 0.75 | 0.75 | 0.907 |
| Middle rank | 0/1 | 0.18 | 0.19 | 0.18 | 0.731 |
| <i>Share correct</i> (middle) | 0/1 | 0.15 | 0.14 | 0.16 | 0.356 |
| Bottom rank | 0/1 | 0.17 | 0.24 | 0.10 | <0.001 |
| <i>Share correct</i> (bottom) | 0/1 | 0.10 | 0.11 | 0.10 | 0.363 |
| <i>N</i> | | 575 | 287 | 288 | 575 |

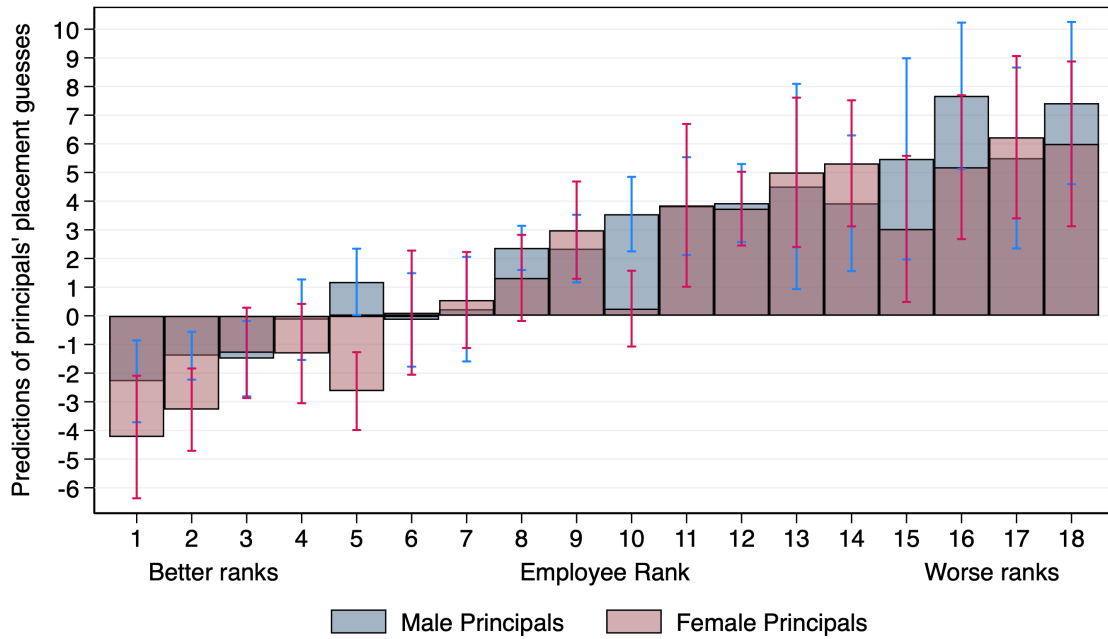
NOTES: The table reports principals' guesses about employee rank beliefs separately by principal gender. The rows labelled "diff. (princ.-emp.)" and "diff. (emp.-princ.)" measure differences between principals' guesses and the actual behaviour of employees assigned to principals of each respective gender. For variables where the difference is "princ.-emp.", positive values indicate that the principals' guess is on average higher compared to the employees' actual choices, while negative values indicate the opposite. Specifically, for the preferred rank the difference is "emp.-princ.", for which a positive value indicate that the principals assessed the employees' to select a better rank compared to the employees' actual choices, and *vice versa* for negative values. Standard deviations for these difference rows are identical to those of the principals' original guesses and are thus omitted for clarity. The last column reports p -values from two-sample t -tests comparing female and male principal groups.

4.2 Principals' confidence guesses

I now examine gender differences among principals in predicting employees' self-confidence, again based on employees' relative performance ranks alone. Given the bounded nature of employees' rank assessments (limited to ranks 1 through 18), higher-performing employees (numerically lower ranks) are mechanically more likely to underestimate their ranks, whereas lower-performing employees (numerically higher ranks) are mechanically more likely to overestimate their ranks. Consequently, self-confidence, measured as self-placement accuracy, tends to shift from negative (underplacement) to positive (overplacement) as the observed rank worsens (relative performance goes down). [Figure 3](#) visualises this pattern: principals typically perceive employees as underconfident (negative values) for the highest ranks (approximately ranks 1–5) and as overconfident (positive values) for the lowest ranks (approximately ranks 8–18).

The figure further illustrates gender differences among principals regarding these perceptions by plotting predicted self-placement accuracy separately for female and male principals. Notably, female principals consistently believe their employees to be more underconfident for higher ranks compared to male principals.¹⁸

Figure 3: Predicted employee selfconfidence by rank and principal gender



NOTES: The figure displays predicted employee placement guesses (selfconfidence measured as self-assessed rank accuracy) by actual (observed) employee rank, separately by principal gender. Bars represent predicted values, and whiskers represent 95% confidence intervals. Predictions are derived from an OLS regression of principals' placement guesses on employee rank, principal gender, their interaction, and control variables including indicators for the employee being top-ranked among the principal's three employees, rank range among employees, principal age, completed years of higher education, field of study indicators (STEM, Languages/Linguistics/Literature), and a principal risk-preference indicator. Standard errors are clustered at the principal level.

[Table 4](#) presents these gender differences in regression format. Principals infer employee self-confidence

¹⁸The regression specification used in [Figure 3](#) closely matches that of column 3 in [Table 4](#), with the exception that the interaction between employee rank and principal gender is explicitly included in the figure.

based on their perceptions of employees’ self-placement accuracy, capturing whether employees are predicted to overestimate (positive values), accurately estimate (zero), or underestimate (negative values) their true rank. The significantly negative coefficient on *Female principal* indicates that female principals consistently predict lower self-confidence (less overplacement or greater underplacement) compared to male principals, even after controlling for observed employee ranks (columns 2–3; $p < 0.01$). For the full sample, the negative coefficient on *Female principal* (-0.98) indicates that female principals predict employee self-confidence approximately 0.21 standard deviations lower than male principals, representing a statistically significant gender difference (two-sided t -test, $p < 0.01$). Among top-performing employees, the negative coefficient (-1.13) indicates that female principals predict employee self-confidence approximately 0.32 standard deviations lower than male principals (two-sided t -test, $p < 0.01$). Conversely, among bottom-half performers (columns 8–9), significant gender differences in principals’ predictions disappear.

Additionally, the positive, and statistically significant coefficients, for *Rank emp.* indicate that principals systematically predict higher-ranked (numerically worse-ranked) employees as more likely to overestimate their ranks. This illustrates a consistent pattern in principals’ predictions of employee self-confidence based purely on observed performance ranks.

Table 4: Principals’ confidence guesses

| | All Ranks | | | Top Half | | | Bottom Half | | |
|-----------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Female principal | | -0.942** (0.340) | -0.980** (0.336) | | -1.134** (0.406) | -1.139** (0.412) | | -0.690 (0.626) | -0.537 (0.645) |
| Rank emp. | 0.553*** (0.043) | 0.558*** (0.043) | 0.586*** (0.049) | 0.618*** (0.067) | 0.616*** (0.067) | 0.716*** (0.081) | 0.530*** (0.107) | 0.538*** (0.105) | 0.478*** (0.107) |
| Employee covariates | | | ✓ | | | ✓ | | | ✓ |
| Principal covariates | | | ✓ | | | ✓ | | | ✓ |
| Mean of dep. variable | 1.79 | 1.79 | 1.79 | -0.47 | -0.47 | -0.47 | 4.54 | 4.54 | 4.54 |
| Observations | 575 | 575 | 575 | 316 | 316 | 316 | 259 | 259 | 259 |
| #Clusters | 192 | 192 | 192 | 182 | 182 | 182 | 165 | 165 | 165 |

Standard errors in parentheses, clustered on principals.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

NOTES: The table shows results from ordinary least squares (OLS) regressions of principals’ predicted employee selfconfidence (employees’ self-placement accuracy) on principal gender and employee rank. Columns 3, 6, and 9 include both employee- and principal-level controls: an indicator for whether the employee is the principal’s top-ranked; the range between the highest- and lowest-ranked employees observed by the principal; the principal’s age; completed years of higher education; indicators for majoring in STEM subjects, and in Languages, Literature, or Linguistics; and an indicator for whether the principal selects the riskier lottery.

4.3 Principals’ gender guesses

Table 5 presents linear probability model (LPM) results analysing how principals predict the gender of their employees after receiving treatment information regarding employee self-placement accuracy. The dependent variable is an indicator of whether the principal guesses the employee to be female. The main independent variables of interest are interactions between the treatment condition (*Reveal*) and observed self-confidence types (underplacement, accurate placement, and overplacement).

Principals significantly associate underplacement (employees who underestimate their ranks) with being female, indicated by the consistently positive and statistically significant coefficients on $Reveal \times Underpl.$ across specifications (columns 1–5). Specifically, principals exposed to self-confidence information are about 18% percentage points more likely to guess an employee is female if the employee underestimated their rank, relative to when no self-assessment information is revealed. Conversely, the negative and statistically significant coefficients on $Reveal \times Overpl.$ indicate that principals associate overplacement (employees who overestimate their ranks) with being male. Principals exposed to self-confidence information are roughly 10% less likely to guess that an employee who overestimates their rank is female. The interaction term for accurate placement is statistically insignificant, suggesting that principals do not strongly associate accurate self-assessment with either gender. Notably, columns 4–5 (top-performing employees) show even stronger associations, where underplacement strongly predicts guessing female gender (with coefficients around 24% p.p.), while overplacement associations diminish in statistical significance. Among bottom-performing employees (columns 6–7), principals continue to associate overplacement with being male (with coefficients around 13% p.p.).¹⁹

Overall, these results suggest principals hold distinct and statistically significant gender stereotypes about employees’ self-confidence, particularly regarding underplacement and overplacement. Importantly, these associations persist even after controlling for the actual observed performance ranks of employees. Thus, principals’ gender predictions reflect perceptions and stereotypes about self-confidence inaccuracies that go beyond simply interpreting relative employee performance. Such persistent stereotypes, if generalised to broader labour-market contexts, could have meaningful implications for workplace interactions, employee evaluations, task allocation, and promotion decisions.

¹⁹Table C.1 in Appendix C show results separated by principal gender. While these results indicate that the gender stereotyping is stronger among women, separate analysis not demonstrated here, does not reveal any statistically significant gender differences in treatment effects.

Table 5: Principals’ gender guesses

| | All Ranks | | | Top Half | | Bottom Half | |
|-------------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Female Principal | | | 0.050 (0.030) | | 0.140** (0.044) | | -0.074 (0.059) |
| <i>Reveal treatment</i> | | | | | | | |
| Reveal × Underpl. | 0.178** (0.060) | 0.179** (0.060) | 0.181** (0.060) | 0.243*** (0.064) | 0.245*** (0.063) | -0.095 (0.157) | -0.105 (0.155) |
| Reveal × Accurate pl. | 0.046 (0.113) | 0.046 (0.112) | 0.040 (0.110) | 0.052 (0.123) | 0.022 (0.118) | 0.155 (0.277) | 0.139 (0.289) |
| Reveal × Overpl. | -0.109** (0.041) | -0.107* (0.042) | -0.107* (0.042) | -0.099 (0.065) | -0.111 (0.064) | -0.133* (0.060) | -0.137* (0.060) |
| Constant | 0.479*** (0.044) | 0.494*** (0.055) | 0.482*** (0.054) | 0.448*** (0.033) | 0.385*** (0.037) | 0.511*** (0.040) | 0.553*** (0.054) |
| Employee Rank | ✓ | ✓ | ✓ | | | | |
| Covariates (principal) | | ✓ | ✓ | | | | |
| Mean of dep. variable | 0.475 | 0.475 | 0.475 | 0.494 | 0.494 | 0.452 | 0.452 |
| Observations | 575 | 575 | 575 | 316 | 316 | 259 | 259 |
| #Clusters | 192 | 192 | 192 | 182 | 182 | 165 | 165 |

Standard errors in parenthesis, clustered on principals.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

NOTES: The table shows results from OLS regressions predicting principals’ guesses of employee female gender. Column (2) and (3) include principal-level controls: principal’s completed years of higher education, indicators for majoring in STEM subjects, Languages, Literature, and Linguistics, and an indicator for choosing the riskier lottery. All specifications control for employee rank. Interaction terms represent principals observing self-placement information (Reveal) combined with the type of self-confidence observed (underplacement, accurate placement, or overplacement).

5 Concluding Remarks

This paper experimentally examines how employers interpret employees’ self-confidence signals, specifically investigating gender differences among managers (principals) in evaluating biased self-assessments of relative performance. The results reveal substantial gender differences in managerial perceptions. Female principals consistently perceive employees as significantly more uncertain and pessimistic about their performance ranks compared to male principals. Particularly among top performers, female principals predict lower self-confidence levels—by approximately 0.2–0.3 standard deviations—compared to their male counterparts.

Moreover, experimental findings demonstrate pronounced gender stereotyping once managers observe employee self-confidence inaccuracies. Principals strongly associate underconfidence predominantly with female employees and overconfidence predominantly with male employees, clearly reflecting entrenched gender biases. If such perceptions and stereotypes extend beyond the laboratory, they could influence workplace dynamics, affecting task allocation, career progression, and compensation. Female managers’ more pessimistic perceptions might inadvertently limit employee opportunities for challenging tasks and advancement, whereas male managers’ overestimations could lead to task misallocation and inefficiencies. Overall, this research provides valuable insights into the nuanced role

managerial gender and perceptions play in shaping workplace dynamics, highlighting implications for gender equality and organisational efficiency.

This study contributes to existing research by highlighting managerial gender as a moderator of how self-confidence signals are interpreted, extending insights from prior work (e.g., [Thoma, 2016](#)). However, this study is not without limitations. The experimental setting, while allowing precise control and identification, might differ from natural employment contexts. Thus, an important avenue for future research would be to examine whether and how these laboratory-induced perceptions generalise to real-world hiring, promotion, and evaluation decisions. Additionally, future studies should explore interventions that might mitigate managerial stereotypes, particularly educational or informational interventions aimed explicitly at reducing gender-biased interpretations of employee behaviours (*see* e.g., [Goldin and Rouse, 2000](#), for an application).

References

- Babcock, L., Recalde, M. P., Vesterlund, L., and Weingart, L. (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, 107(3):714–747.
- Bandiera, O., Parekh, N., Petrongolo, B., and Rao, M. (2022). Men are from mars, and women too: a bayesian meta-analysis of overconfidence experiments. *Economica*, 89:S38–S70.
- Barber, B. M. and Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1):261–292.
- Barron, K., Dittmann, R., Gehrig, S., and Schweighofer-Kodritsch, S. (2025). Explicit and implicit belief-based gender discrimination: A hiring experiment. *Management Science*, 71(2):1600–1622.
- Bengtsson, C., Persson, M., and Willenhag, P. (2005). Gender and overconfidence. *Economics letters*, 86(2):199–203.
- Bohnet, I., Van Geen, A., and Bazerman, M. H. (2016). When performance trumps gender bias: Joint versus separate evaluation. *Management Science*, 62(5):1225–1234.
- Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794.
- Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, 109(3):739–73.
- Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.
- Clark, J. and Friesen, L. (2009). Overconfidence in forecasts of own performance: An experimental study. *The Economic Journal*, 119(534):229–251.
- Coffman, K. B., Exley, C. L., and Niederle, M. (2021). The role of beliefs in driving gender discrimination. *Management Science*, 67(6):3551–3569.
- Exley, C. L. and Kessler, J. B. (2022). The gender gap in self-promotion*. *The Quarterly Journal of Economics*, 137(3):1345–1381.
- Goldin, C. and Rouse, C. (2000). Orchestrating impartiality: The impact of “blind” auditions on female musicians. *American economic review*, 90(4):715–741.
- Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1):114–125.
- Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior*, 32:113–135.
- Lénárd, T., Horn, D., and Kiss, H. J. (2024). Competition, confidence and gender: Shifting the focus from the overconfident to the realistic. *Journal of Economic Psychology*, 104:102746.

- Malmendier, U. and Tate, G. (2008). Who makes acquisitions? ceo overconfidence and the market's reaction. *Journal of financial Economics*, 89(1):20–43.
- Moore, D. A. and Cain, D. M. (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior and Human Decision Processes*, 103(2):197–213.
- Reuben, E., Sapienza, P., and Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12):4403–4408.
- Sarsons, H. (2017). Interpreting signals in the labor market: evidence from medical referrals. *Job market paper*, pages 141–145.
- Sarsons, H., Gërxhani, K., Reuben, E., and Schram, A. (2021). Gender differences in recognition for group work. *Journal of Political Economy*, 129(1):101–147.
- Thoma, C. (2016). Under-versus overconfidence: an experiment on how others perceive a biased self-assessment. *Experimental Economics*, 19:218–239.

APPENDICES

A Additional questions in the Principals' Survey

- The following two additional questions were included in the principals' survey, before the gender-question that was originally recorded as question 8.

6. **Principal's payment option (real-choice):** Principals can select their payment principal (range: Employee 1–Employee 3, Yes/No).

“Instead of being paid by 1/3 from each of your Employees. Would you like to be paid entirely by Employee $Q5_{E1/E2/E3}$?”

Yes ☐ No ☐

7. **Performance improvements:** Principals guess, who, if any, that they think will score higher in the next experimental round. (range: none; one; or multiple)

“Which combination of your Employees, if any do you think is going to score a higher score in the next round?”

Employee 1 ☐ Employee 2 ☐ Employee 3 ☐ None ☐

B Balance Testing and Descriptive Statistics

Table B.1: Covariates and balance test - by principals' treatment conditions

| | Joint F -test | NoREVEAL | | | REVEAL | | | Pairwise t test |
|------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-------------------|
| | | Total | NoFeedback | Feedback | Total | NoFeedback | Feedback | |
| | $p(F)$ | mean/(sd) | | | mean/(sd) | | | p -value |
| | [(1):(4)] | (A) | (1) | (2) | (B) | (3) | (4) | (A)-(B) |
| RET Score | 0.245 | 16.29 (4.16) | 16.62 (3.96) | 15.96 (4.37) | 16.32 (3.44) | 15.60 (3.56) | 17.04 (3.19) | 0.955 |
| Age | 0.003 | 23.84 (2.96) | 23.90 (3.20) | 23.79 (2.73) | 24.94 (4.63) | 26.21 (5.74) | 23.67 (2.65) | 0.053 |
| Finished study years | 0.989 | 3.34 (1.80) | 3.31 (1.79) | 3.38 (1.83) | 3.32 (1.70) | 3.38 (1.81) | 3.27 (1.59) | 0.934 |
| Majors in STEM-subject | 0.392 | 0.19 | 0.25 | 0.12 | 0.23 | 0.21 | 0.25 | 0.480 |
| Majors in LLL-subject | 0.377 | 0.03 | 0.00 | 0.06 | 0.05 | 0.06 | 0.04 | 0.473 |
| Languages spoken | 0.728 | 3.18 (0.92) | 3.15 (0.97) | 3.21 (0.87) | 3.27 (0.98) | 3.35 (1.00) | 3.19 (0.96) | 0.494 |
| Economic situation | 0.748 | 4.82 (1.86) | 4.88 (2.14) | 4.77 (1.55) | 5.08 (1.83) | 5.00 (2.15) | 5.17 (1.45) | 0.329 |
| Selects riskier lottery | 0.362 | 0.62 | 0.58 | 0.67 | 0.69 | 0.75 | 0.62 | 0.365 |
| <i>Comprehension (proxy)</i> | | | | | | | | |
| Correct 1st attempt | 0.392 | 4.04 (1.16) | 4.23 (1.10) | 3.85 (1.20) | 4.04 (1.08) | 3.98 (1.04) | 4.10 (1.12) | 1.000 |
| N | 192 | 96 | 48 | 48 | 96 | 48 | 48 | 192 |

Table B.2: Summary statistics of employees - performance and rank beliefs, by gender: by principals' treatment

| | | All | | | | NoReveal | | | | Reveal | | | |
|----------------------------|--------|----------------|----------------|----------------|-----------------------|----------------|----------------|----------------|-----------------------|----------------|----------------|----------------|------------------------|
| | Scale | (1) All | (2) Male | (3) Female | (4) <i>p</i> -val. | (5) All | (6) Male | (7) Female | (8) <i>p</i> -val. | (9) All | (10) Male | (11) Female | (12) <i>p</i> -val. |
| <i>Rank beliefs</i> | | | | | | | | | | | | | |
| No. of ranks | 1-18 | 5.29 (3.06) | 5.34 (2.83) | 5.25 (3.28) | 0.722 | 5.37 (3.24) | 5.38 (2.81) | 5.36 (3.63) | 0.962 | 5.22 (2.88) | 5.30 (2.86) | 5.13 (2.90) | 0.630 |
| Actual rank | 1-18 | 8.89 (5.28) | 8.57 (5.58) | 9.21 (4.95) | 0.147 | 9.04 (5.28) | 8.96 (5.76) | 9.12 (4.76) | 0.797 | 8.75 (5.28) | 8.19 (5.38) | 9.31 (5.14) | 0.073 |
| % of invested | 0-1 | 0.08 (0.14) | 0.09 (0.13) | 0.08 (0.16) | 0.521 | 0.08 (0.15) | 0.08 (0.13) | 0.08 (0.16) | 0.793 | 0.08 (0.14) | 0.09 (0.13) | 0.07 (0.15) | 0.231 |
| Preferred rank | 1-18 | 6.25 (4.33) | 5.53 (4.10) | 6.98 (4.44) | 0.000 | 5.97 (4.41) | 5.26 (3.93) | 6.69 (4.75) | 0.006 | 6.53 (4.25) | 5.81 (4.26) | 7.27 (4.11) | 0.003 |
| % of invested ¹ | 0-1 | 0.37 (0.17) | 0.35 (0.14) | 0.38 (0.19) | 0.014 | 0.37 (0.18) | 0.34 (0.13) | 0.39 (0.21) | 0.010 | 0.37 (0.16) | 0.36 (0.16) | 0.38 (0.17) | 0.398 |
| <i>Selfconfidence</i> | | | | | | | | | | | | | |
| Placement ² | -16-16 | 2.64 (5.88) | 3.04 (5.68) | 2.24 (6.06) | 0.101 | 3.07 (5.69) | 3.70 (5.48) | 2.43 (5.85) | 0.059 | 2.22 (6.04) | 2.39 (5.81) | 2.04 (6.28) | 0.629 |
| <i>Underplac.</i> | 0/1 | 0.29 | 0.26 | 0.33 | 0.068 | 0.26 | 0.22 | 0.30 | 0.098 | 0.32 | 0.30 | 0.35 | 0.335 |
| <i>Accurate plac.</i> | 0/1 | 0.08 | 0.07 | 0.08 | 0.727 | 0.08 | 0.06 | 0.10 | 0.269 | 0.07 | 0.08 | 0.06 | 0.518 |
| <i>Overplac.</i> | 0/1 | 0.63 | 0.67 | 0.59 | 0.056 | 0.66 | 0.72 | 0.60 | 0.030 | 0.60 | 0.62 | 0.59 | 0.564 |
| <i>N</i> | | 575 | 289 | 286 | 575 | 287 | 144 | 143 | 287 | 288 | 145 | 143 | 288 |

1) The investment allocation for *the most preferred rank* must be ≤ 2 ECUs, which gives a lower range of approximately 0.11 ($= 2/19$).

2) Placement is measured as *true rank* - *preferred rank*, so that a negative (positive) number implies the underplacement (overplacement) of agent's rank and a 0 signifies an accurate placement regarding the most preferred rank.

Table B.3: Balance tests of principals' baseline guesses by treatment condition, by gender

| | Scale | All | | | | Male | | | | Female | | | |
|---|-------|----------------|-----------------|----------------|-----------------------|----------------|-----------------|----------------|-----------------------|----------------|------------------|----------------|------------------------|
| | | All (1) | NoReveal (2) | Reveal (3) | <i>p</i> -val. (4) | All (5) | NoReveal (6) | Reveal (7) | <i>p</i> -val. (8) | All (9) | NoReveal (10) | Reveal (11) | <i>p</i> -val. (12) |
| <i>Principals' guesses on employees</i> | | | | | | | | | | | | | |
| Rank diversification | 1-18 | 6.49 (4.54) | 6.43 (4.12) | 6.55 (4.93) | 0.759 | 5.65 (3.75) | 5.55 (3.43) | 5.75 (4.04) | 0.644 | 7.32 (5.09) | 7.31 (4.55) | 7.34 (5.58) | 0.954 |
| Preferred rank | 1-18 | 7.11 (4.41) | 7.08 (4.42) | 7.13 (4.41) | 0.903 | 6.52 (4.15) | 6.51 (4.11) | 6.52 (4.19) | 0.983 | 7.69 (4.59) | 7.65 (4.65) | 7.74 (4.54) | 0.878 |
| Belief weight—preferred rank | 2-19 | 8.14 (4.05) | 8.09 (3.70) | 8.18 (4.38) | 0.790 | 8.24 (3.74) | 8.28 (3.64) | 8.20 (3.84) | 0.859 | 8.04 (4.35) | 7.91 (3.76) | 8.17 (4.88) | 0.617 |
| Belief weight—actual rank | 0-19 | 3.41 (3.29) | 3.43 (3.23) | 3.39 (3.35) | 0.905 | 3.68 (3.52) | 3.69 (3.66) | 3.66 (3.40) | 0.938 | 3.14 (3.02) | 3.16 (2.74) | 3.12 (3.29) | 0.922 |
| <i>N</i> | | 575 | 287 | 288 | 575 | 287 | 143 | 144 | 287 | 288 | 144 | 144 | 288 |

Standard deviations in parentheses.

p-values are from *t*-tests of mean differences.

NOTES: The table reports tests for differences in the means of principals' baseline perceptions, by gender, across the two principal treatment conditions, *NoReveal* and *Reveal*. The variables tested include principals' guesses about employees' rank diversification (number of ranks selected), the employees' most preferred rank, the strength of employees' beliefs (belief weights) in their most preferred rank, and the strength of employees' beliefs in their actual rank. Means, standard deviations (in parentheses), and *p*-values from two-sided *t*-tests for differences between treatments are provided. These balance tests verify whether principals' initial perceptions systematically differ by treatment condition, an important check for ensuring internal validity of subsequent analyses. Number of observations (*N*) corresponds to the number of employee observations.

Table B.4: All principals guesses across principal treatment, by employee treatment

| | Scale | NoFeedback (All) | | | | Feedback (All) | | | |
|---|-------|------------------|-----------------|----------------|-----------------------|----------------|-----------------|----------------|-----------------------|
| | | All (1) | NoReveal (2) | Reveal (3) | <i>p</i> -val. (4) | All (5) | NoReveal (6) | Reveal (7) | <i>p</i> -val. (8) |
| <i>Principals' guesses on employees</i> | | | | | | | | | |
| Rank diversification | 1-18 | 6.56 (4.57) | 6.71 (4.06) | 6.40 (5.03) | 0.571 | 6.42 (4.52) | 6.15 (4.17) | 6.69 (4.84) | 0.312 |
| Preferred rank | 1-18 | 7.32 (4.26) | 7.17 (4.23) | 7.46 (4.30) | 0.571 | 6.90 (4.55) | 6.99 (4.61) | 6.80 (4.51) | 0.718 |
| Belief weight on preferred rank | 2-19 | 8.31 (4.04) | 7.85 (3.38) | 8.77 (4.57) | 0.052 | 7.97 (4.06) | 8.34 (3.98) | 7.60 (4.12) | 0.120 |
| Belief weight on actual rank | 0-19 | 3.58 (3.17) | 3.81 (3.24) | 3.35 (3.09) | 0.213 | 3.24 (3.40) | 3.03 (3.19) | 3.44 (3.60) | 0.317 |
| <i>N</i> | | 288 | 144 | 144 | 288 | 287 | 143 | 144 | 287 |

Table B.5: Male principals guesses across principal treatment, by employee treatment

| | Scale | NoFeedback (Male) | | | | Feedback (Male) | | | |
|---|-------|-------------------|-----------------|----------------|-----------------------|-----------------|-----------------|----------------|-----------------------|
| | | All (1) | NoReveal (2) | Reveal (3) | <i>p</i> -val. (4) | All (5) | NoReveal (6) | Reveal (7) | <i>p</i> -val. (8) |
| <i>Principals' guesses on employees</i> | | | | | | | | | |
| Rank diversification | 1-18 | 5.66 (3.58) | 5.62 (2.94) | 5.69 (4.14) | 0.908 | 5.64 (3.92) | 5.46 (3.88) | 5.81 (3.97) | 0.605 |
| Preferred rank | 1-18 | 6.88 (4.08) | 6.49 (3.70) | 7.28 (4.42) | 0.246 | 6.15 (4.19) | 6.54 (4.52) | 5.76 (3.84) | 0.273 |
| Belief weight on preferred rank | 2-19 | 8.67 (3.75) | 8.49 (3.39) | 8.86 (4.10) | 0.550 | 7.80 (3.68) | 8.07 (3.89) | 7.54 (3.48) | 0.392 |
| Belief weight on actual rank | 0-19 | 3.81 (3.50) | 3.96 (3.63) | 3.67 (3.37) | 0.618 | 3.54 (3.56) | 3.42 (3.69) | 3.65 (3.45) | 0.700 |
| <i>N</i> | | 144 | 72 | 72 | 144 | 143 | 71 | 72 | 143 |

Table B.6: Female principals guesses across principal treatment, by employee treatment

| | Scale | NoFeedback (Female) | | | | Feedback (Female) | | | |
|---|-------|---------------------|----------------|----------------|--------|-------------------|----------------|----------------|--------|
| | | All | NoReveal | Reveal | p-val. | All | NoReveal | Reveal | p-val. |
| | | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| <i>Principals' guesses on employees</i> | | | | | | | | | |
| Rank diversification | 1-18 | 7.45 (5.24) | 7.79 (4.71) | 7.11 (5.73) | 0.438 | 7.19 (4.94) | 6.82 (4.36) | 7.57 (5.47) | 0.364 |
| Preferred rank | 1-18 | 7.75 (4.40) | 7.86 (4.62) | 7.64 (4.19) | 0.763 | 7.64 (4.78) | 7.44 (4.69) | 7.83 (4.90) | 0.627 |
| Belief weight on preferred rank | 2-19 | 7.94 (4.29) | 7.21 (3.28) | 8.68 (5.03) | 0.039 | 8.13 (4.42) | 8.61 (4.09) | 7.65 (4.70) | 0.194 |
| Belief weight on actual rank | 0-19 | 3.35 (2.80) | 3.67 (2.81) | 3.03 (2.76) | 0.171 | 2.94 (3.22) | 2.65 (2.58) | 3.22 (3.76) | 0.291 |
| <i>N</i> | | 144 | 72 | 72 | 144 | 144 | 72 | 72 | 144 |

NOTES: The tables report means and standard deviations (in parentheses) of principals' baseline guesses regarding employees' rank diversification, preferred rank, belief weight assigned to the preferred rank, and belief weight assigned to the actual rank, separated by employee treatment conditions (*NoFeedback* vs. *Feedback*). *P*-values from two-sided *t*-tests for differences between the principal treatment conditions (*NoReveal* vs. *Reveal*) within each employee treatment are provided. Number of observations (*N*) refers to employee observations. These test the internal validity of treatment randomisation across employee-side treatment variations.

C Robustness

Table C.1: (OLS) Principals' gender guesses: by principal gender

| | Male Principals | | Female Principals | |
|------------------------------|---------------------|---------------------|---------------------|---------------------|
| | (1) | (2) | (3) | (4) |
| Reveal \times Underpl. | 0.169 (0.088) | 0.161 (0.089) | 0.201* (0.078) | 0.207* (0.079) |
| Reveal \times Accurate pl. | -0.193 (0.154) | -0.184 (0.157) | 0.158 (0.139) | 0.170 (0.138) |
| Reveal \times Overpl. | -0.080 (0.063) | -0.094 (0.063) | -0.142** (0.054) | -0.140* (0.055) |
| Constant | 0.394*** (0.064) | 0.454*** (0.073) | 0.561*** (0.062) | 0.540*** (0.089) |
| Employee Rank | ✓ | ✓ | ✓ | ✓ |
| Covariates (principal) | | ✓ | | ✓ |
| Mean of dep. variable | 0.456 | 0.456 | 0.493 | 0.493 |
| Observations | 287 | 287 | 288 | 288 |
| #Clusters | 96 | 96 | 96 | 96 |

Standard errors in parenthesis, clustered on principals.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

D Experimental Instruction (translated to English)