Social Preferences and the Informativeness of Subjective Performance Evaluations*

David J. Kusterer† Dirk Sliwka‡

October 26, 2023

Abstract

We study biases and the informativeness of subjective performance evaluations in an MTurk experiment, testing the implications of a standard formal framework of rational subjective evaluations. In the experiment, subjects in the role of workers perform a real effort task. Subjects in the role of supervisors observe samples of the workers' output and assess their performance. We conduct 6 experimental treatments varying (i) whether workers' pay depends on the performance evaluation, (ii) whether supervisors are paid for the accuracy of their evaluations, and (iii) the precision of the information available to supervisors. Moreover, we use the exogenous assignment of supervisors to workers to investigate the association between supervisors' social preferences and their rating quality. In line with the model of optimal evaluations, we find that ratings are more lenient and less informative when they determine bonus payments. Rewards for accuracy reduce leniency and can enhance informativeness. When supervisors have access to more detailed performance information, their ratings vary more with the performance signal and become more informative. Contrary to expectations, we do not find that more prosocial supervisors are systematically more lenient when their ratings affect worker's payoffs. Instead, they are more diligent in their rating behavior, resulting in more accurate and informative performance evaluations.

Keywords: Subjective performance evaluation, bias, bonuses, differentiation, social preferences

*This study has been pre-registered in the AEA RCT Registry (RCT ID: AEARCTR-0005020). Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1 – 390838866 is gratefully acknowledged. We would like to thank Nicolas Fugger, Evelyn Intan, Felix Kölle, Johannes Mans, Frank Moers, Patrick Schmitz, Marina Schröder, participants of the 2021 annual meeting of the German Association for Experimental Economics in Magdeburg, the Colloquium on Personnel Economics 2022 in Aarhus, the Maastricht Behavioral Economic Policy Symposium 2022, the European ESA meeting 2022 in Bologna, the Meeting of the German Economic Association 2022 in Basel, and the Symposium on the Economic Analysis of the Firm 2023 (GEABA) in Paderborn as well as seminar participants at the Rotterdam School of Management, the University of Leicester and City, University of London for helpful comments and discussions. Moreover, we would like to thank Mary Wack and Niklas Wagner for excellent research assistance.

†University of Cologne. Email: kusterer@uni-koeln.de

‡University of Cologne, CEPR, CESifo, and IZA. Email: sliwka@wiso.uni-koeln.de

1 Introduction

In many jobs, an employee's performance cannot fully be assessed by objective key figures. Hence, it is common that firms ask supervisors to subjectively rate the performance of their subordinates. However, these evaluations are known to be be "biased", and there are systematic deviations of the distribution of subjective assessments from the underlying distribution of performance.

For instance, ratings tend to be overly lenient and compressed (see e.g. Murphy and Cleveland, 1995; Prendergast and Topel, 1996; Prendergast, 1999). This means that average ratings exceed average performance, and the variance of ratings does not fully reflect the the variance of the underlying performance outcomes. These "biases" can be caused by various mechanisms. For one, supervisors, who carry out the performance evaluations but are not owners of the organization, may have *preferences* that are not aligned with the employers'. Supervisors thus may intentionally deviate from accurate assessments, especially if they exhibit social preferences towards the employee and need to trade off the employer's interests against the assessed employee's. Inaccurate ratings can also result from insufficient *information.* Even when supervisors value the accuracy of their ratings, they may have incomplete information about the employee's true performance. Finally, *cognitive limitations* could cause observed deviations between true performance and performance assessments. Supervisors may not process the available information in a fully rational manner, potentially leading to inaccurate evaluations.

Firms use subjective performance evaluations for multiple purposes (see e.g. Landy and Farr, 1983; Arvey and Murphy, 1998; Prendergast, 1999). For instance, they are used to allocate individual bonuses in incentive schemes when objective performance measures are unavailable. Additionally, evaluations are used for personnel decisions such as promotions or terminations. Particularly in these cases, biases can be costly if they reduce the informativeness of ratings and then lead to distorted personnel decision with long-term consequences.

In this paper, we build on a standard framework (Prendergast and Topel, 1996) to formally model subjective performance evaluations by a rational decision-maker, and test its implications experimentally.[1] In the framework, supervisors rate an agent's performance based on the observation of noisy performance signals. Supervisors trade off a preference for rating accuracy – which implies the application of Bayes' rule for an optimal rating given the noisy information – against potential social preferences towards the assessed worker, which imply a tendency for more lenient and consequently, less accurate ratings.

[1]The framework is e.g. extended in Golman and Bhatia (2012) to account for asymmetric reactions to good versus bad ratings, and applied in Manthei and Sliwka (2019) to study the interplay between multitasking and subjective performance evaluations and in Kampkötter and Sliwka (2018) to study the allocation of bonuses in teams. The related literature on "muddled information" (e.g. Frankel and Kartik, 2019, 2021; Ball, 2022) is also based on this framework.

To test the model's predictions, we conducted an online experiment involving 780 subjects on Amazon MTurk, a crowdsourced labor platform. Participants worked on a real-effort task of entering text from hard-to-read images, similar to "captchas". Subsequently, another group of subjects in the role of supervisors evaluated each participant's performance, defined as the percentage of correctly entered images. Supervisors had limited information on workers' performance, as they could only view a randomly selected subset of the workers' performance outcomes. We implemented six different treatments, varying (i) whether workers' pay was tied to the rating, (ii) whether supervisors' pay was tied to accuracy, and (iii) whether supervisors observed a smaller or larger subset of the workers' performance outcomes.

We also use the random assignment of supervisors to agents to examine the association between supervisors' social preferences and their rating behavior. We measure each supervisor's social preferences using the incentivized Social Value Orientation (SVO) measure (Murphy et al., 2011). The SVO measure comprises several dictator games where subjects allocate money between themselves and a randomly matched worker. The choices reveal the weight a decision-maker places on the receiver's payoff relative to their own. Notably, we therefore are not measuring favoritism as in the original PT framework, i.e., differential social preferences towards specific workers, but general social preferences towards the worker population.

We derive hypotheses for each treatment variation based on the framework. In particular, we consider treatment effects on (i) rating leniency and (ii) rating quality. To assess the latter, we first examine the *rating error*, defined as the (squared) deviation between the rating and true performance, as a measure of rating quality when the ratings are taken "at face value". Crucially, we also evaluate the *informativeness* of the ratings, i.e., their usefulness for a rational decision-maker aiming to predict true performance for personnel decisions, while being aware of potential biases.

We find that the treatment-induced patterns are mostly well-organized by the formal model. First, for the same performance signals, ratings are significantly higher when workers' pay is tied to the ratings.[2] This suggests that supervisors internalize the effect of their ratings on a worker's well-being, even in an anonymous experiment without future interaction. Second, rewarding supervisors for rating accuracy mitigates the rating leniency induced by worker bonuses. Lastly, when supervisors have access to a larger number of signals, their ratings vary more with observed performance signals.

Consistent with the model's predictions, performance pay for workers tends to reduce

[2]In their standard textbook on performance appraisals, Murphy and Cleveland (1995), for instance, conjectured that "As PA [Performance Appraisal] is more and more closely linked to important rewards, we expect that the pressure to give high ratings will become even more severe" (p. 344). In a meta-analysis, Jawahar and Williams (1997) find that ratings obtained for pay raises or promotions are more lenient than ratings obtained for research or feedback purposes.

the informativeness of the ratings, while rewarding supervisors for accuracy and providing more performance signals increases informativeness. Hence, the results underscore the inherent tension between different uses of performance ratings: when ratings are used to reward employees, their value to assess employee performance accurately diminishes.

In one domain, however, the empirical results deviate from our initial hypothesis. We predicted that supervisors with stronger social preferences would provide more lenient evaluations, especially when the rating determines the worker's bonus. The reason for the hypothesis is simple: A supervisor who cares more for a worker's payoffs would give these payoffs more weight in her utility function, leading to upwardly distorted ratings. This distortion would increase the rating error and potentially reduce the informativeness of the ratings if the size of this distortion varies between supervisors and is hard to predict. Contrary to this hypothesis, we found no evidence that supervisors with stronger social preferences provide more lenient ratings under performance pay than those with weaker social preferences. Exploring reasons for this observation, we find that stronger social preferences correlate with more diligent rating behavior. For instance, more prosocial supervisors spend significantly more time on the rating. In fact, they even provide more accurate and informative ratings on average. Hence, our initial view that social preferences only affect ratings by increasing in the weight of the worker's payoffs in the supervisor's utility function was too narrow. Social preferences also influence the diligence applied to the rating task, and this effect effect seems to outweigh the former, resulting in higher overall rating quality.

We contribute to the literature on subjective performance evaluations in several respects. For one, our results complement research in psychology on performance evaluations (see e.g. Rynes et al. (2005); Schleicher et al. (2019) for surveys on this literature). While this literature has documented biases in evaluations, we study the extent to which observed patterns can be organized by a formal framework of rational decision-making and analyze the informativeness of ratings, i.e. their usefulness for predicting actual performance. The experimental literature on subjective performance evaluations in behavioral economics and accounting has mostly focused on the effects of subjective assessments on the behavior of the evaluated workers, while we focus more on the determinants of the ratings per se as well as their informativeness. Berger et al. (2013) show that imposing a forced distribution, i.e. forcing evaluators to differentiate, raises worker performance when these workers work separately. Sebald and Walzl (2014) study reactions to subjective performance evaluations and find that workers negatively reciprocate low ratings even when these ratings do not affect their payoffs. Bellemare and Sebald (2019) show that these reactions depend on the worker's over- and underconfidence.[3]

[3]In a supplementary analysis (see Appendix A.3), we also find that workers punish low ratings and reward high ratings (compared to actual performance, as opposed to worker beliefs as in the aforementioned studies), adding evidence to the claim that another reason for lenient ratings is the supervisor's fear of an

A tool to counteract biased ratings recently used by organizations are calibration committees, where groups of supervisors assign or revise ratings proposed by an employee's direct supervisor. Using data from a large multinational organization, Demeré et al. (2019) provide evidence that calibration committees tend to reduce leniency. In a lab experiment, Ockenfels et al. (2020) find that performance evaluations by multiple raters provide more accurate ratings.[4] Grabner et al. (2020) show evidence in line with the idea that calibration committees discipline supervisors and thus tend to reduce leniency and rating compression.[5] Our results suggest that additional incentives for accuracy are particularly needed when a monetary bonus for the worker is tied to the rating. Without material consequences for the worker, rating leniency is less prevalent, and calibration committees may be less beneficial under such circumstances.

With respect to the noisiness of the performance signal, there is prior evidence from non-incentivized vignette studies showing that less precise signals cause higher rating compression, with ambiguous effects on rating leniency (see e.g. Bol and Smith, 2011 and Bol et al., 2016).[6] Our results also provide clear evidence for more compression when the signal is more noisy, but we find no sizable effects on rating leniency.

Our paper is also related to the growing experimental literature on the effects of performance feedback (see Villeval (2020) for a recent survey). While this literature studies worker's reactions to different forms of feedback information, we investigate a setting where performance information is assessed by a supervisor and focus on the supervisor's rating behavior.

The paper proceeds as follows. Section 2 presents the basic model and derives hypotheses. Section 3 introduces the experimental design. Section 4 presents the experimental results and Section 4.6 concludes.

2 A Simple Model

2.1 The Setup

Consider the following simple framework which builds on Prendergast and Topel (1996, PT) to model subjective performance evaluations. A supervisor evaluates the performance

adverse reaction by the worker (see Golman and Bhatia, 2012 or Ockenfels et al., 2015).

[4]They compare evaluations conducted by a group of supervisors who receive independent signals to ratings by a single supervisor who receives multiple signals and find that in both cases, evaluations are less compressed than ratings by a single supervisor who receives one signal.

[5]For instance, they find that supervisors who give inflated ratings are penalized by receiving lower performance evaluations themselves, while supervisors who give less compressed ratings are more likely to receive a promotion.

[6]In the related setting of feedback provision on online markets, Rice (2012) and Bolton et al. (2019) experimentally study the effect of uncertainty about the cause of quality deficiencies on feedback giving and find that it increases rating leniency and compression.

of an agent. The supervisor observes a vector s of $i = 1, .., n$ noisy performance signals $s_i = a + \varepsilon_i$ where $a \sim N\left(m, \sigma_a^2\right)$ is the agent's true performance and the $\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$ are noise terms. The supervisor has to determine a performance rating r. The agent receives a fixed payment of α and she may receive a bonus $\beta \cdot r$ that depends on the rating. In addition, the agent may obtain a psychological benefit $b \cdot r$ from a rating (i.e. $b > 0$) such that her utility is

$$\alpha + (\beta + b) \cdot r.$$

The supervisor may have social preferences towards the agent and therefore, her utility may be affected by the well-being of the agent. As in PT, the supervisor faces material incentives to provide accurate ratings. However, we allow for the possibility that she also intrinsically cares for the accuracy of her rating. The supervisor's utility function is thus

$$\eta \cdot (\alpha + (\beta + b) \cdot r) - \frac{\gamma + \lambda}{2} E\left[(r - a)^2 \middle| s\right],$$

where η measures the supervisor's social preferences, γ are the supervisor's intrinsic preferences for a small rating error, and λ determines her material incentives to provide ratings that are aligned with true performance.[7]

Supervisors differ in their social preferences. We assume that η follows a normal distribution $\eta \sim N\left(m_\eta, \sigma_\eta^2\right)$, where we assume that $m_\eta > 0$ as the literature on social preferences has systematically shown that a majority of individuals tends to be prosocial (see Fehr and Charness (2023) for a recent survey).[8]

Note that we interpret the weight η the supervisor places on the agent's payoff slightly differently than PT. While in the PT framework, the weight captures favoritism towards specific individuals (and thus for the same supervisor, weights can differ between agents), we study general social preferences towards others in the worker population.[9]

In our treatments, we vary whether the agent receives a bonus $\beta > 0$, whether the supervisor is rewarded for accuracy $\lambda > 0$, and whether the supervisor receives one or four performance signals.

2.2 Optimal Evaluations and Hypotheses

First, note that the signal average $\overline{s} = \frac{1}{n} \sum_{i=1}^{n} s_i$ is a sufficient statistic for estimating a, i.e. $E\left[(r - a)^2 \middle| s\right] = E\left[(r - a)^2 \middle| \overline{s}\right]$. The supervisor's decision problem when choosing the

[7] In PT, incentives for accuracy are purely determined by extrinsic rewards as the employer penalizes the supervisor for deviations between her assessments and an assessment based on the firm's information.

[8] The assumption that η follows a normal distribution is not required for Proposition 1 and Hypotheses 1 to 4, but we use it in the proof of Proposition 2 leading to the informativeness results.

[9] A more general model would encompass both factors, i.e. general social preferences as well as personal favoritism. But as interaction in our experiment is anonymous, supervisors cannot exhibit favoritism towards specific individuals in our experimental setting.

optimal rating is

$$\max_r \eta \cdot (\alpha + (\beta + b) \cdot r) - \frac{\gamma + \lambda}{2} E\left[(r-a)^2 \middle| \bar{s}\right].$$

As $E\left[(r-a)^2 \middle| \bar{s}\right] = V[r-a|\bar{s}] + (E[r-a|\bar{s}])^2$ we have[10]

$$E\left[(r-a)^2 \middle| \bar{s}\right] = \frac{\sigma_a^2 \sigma_\varepsilon^2}{n\sigma_a^2 + \sigma_\varepsilon^2} + \left(r - m - \frac{n\sigma_a^2}{n\sigma_a^2 + \sigma_\varepsilon^2}(\bar{s} - m)\right)^2. \quad (1)$$

The first order condition of the supervisor's optimization problem is

$$\eta(\beta + b) - (\gamma + \lambda)\left(r - m - \frac{n\sigma_a^2}{n\sigma_a^2 + \sigma_\varepsilon^2}(\bar{s} - m)\right) = 0$$

from where we obtain the following result:

Proposition 1. *After having observed performance signal $\bar{s}$ the supervisor reports*

$$r(\bar{s}) \quad = \quad \frac{\sigma_\varepsilon^2}{n\sigma_a^2 + \sigma_\varepsilon^2} \cdot m + \frac{n\sigma_a^2}{n\sigma_a^2 + \sigma_\varepsilon^2} \cdot \bar{s} + \frac{\eta(\beta + b)}{\gamma + \lambda}. \quad (2)$$

The first two terms are equal to the conditional expectation of performance given the observed signal average $\bar{s}$. The third term captures the bias induced by social preferences.

Proposition 1 has several implications that we test in the experiment. In particular, we analyze *rating leniency*, defined as the difference between the rating and the conditional expectation of performance given the signal average $\bar{s}$. This corresponds to $\frac{\eta(\beta+b)}{\gamma+\lambda}$ in $r(\bar{s})$ in our model. We also analyze *rating differentiation*, defined as the slope of the rating function $\frac{\partial r}{\partial \bar{s}}$ as our measure of rating compression. The more differentiated the ratings are, the lower the compression.

The first prediction concerns the role of a bonus tied to performance ratings:

Hypothesis 1: *Bonus payments ($\beta > 0$) lead to higher rating leniency.*

The reason for this result is straightforward: When agents receive a bonus tied to the performance rating, supervisors with social preferences will internalize the effect on the agents' well-being, thus shifting ratings upwards. However, due to the additive separability of the supervisor's utility, this internalization does not affect the marginal effect of the observed signal on reported performance, i.e., the slope of the rating function.[11]

[10] Note that $V[r-a|\bar{s}] = V[a|\bar{s}] = V[a] - \frac{(Cov[a,\bar{s}])^2}{V[\bar{s}]}$.

[11] The separability assumption also excludes income effects, which otherwise could have produced a countervailing effect of higher bonuses, as the supervisor needs to be less lenient when aiming at giving the agent a specific bonus amount. However, several results from the experimental literature on social preferences suggests that income effects are likely to be weak. For instance, Charness and Rabin (2002) and Engelmann and Strobel (2004) find substantial evidence for what the former call "social-welfare preferences" and the latter "efficiency concerns", which predict that subjects would be willing to give up more of their own payoff if the payoff sacrifice comes with a stronger return for the other player.

The next prediction describes the effect of a reward for the supervisor for rating accuracy:

Hypothesis 2: *A reward for accuracy ($\lambda > 0$) reduces rating leniency. This reduction in leniency will be larger when the agent receives performance pay ($\beta > 0$).*

Introducing a reward for accuracy increases the cost of rating leniency. As leniency becomes more costly for the supervisor, accuracy pay results in less lenient ratings. Moreover, since ratings are generally more lenient when there is performance pay, the reduction in leniency is more pronounced in this case. To illustrate, consider a situation where the agent's intrinsic private benefit b from higher ratings is small and there is no bonus. In this case, there would be little rating leniency in the first place, and a supervisor would prefer accurate ratings even without an accuracy reward. However, when performance pay is involved, the urge to deviate from accurate ratings increases, making material incentives for accuracy more important.

We also vary the number of signals observed by the supervisor in the experiment:[12]

Hypothesis 3: *If the supervisor observes more signals, rating differentiation increases. That is, the slope of the rating function increases, $\frac{\partial^2 r}{\partial \bar{s} \partial n} > 0$ and its intercept decreases, $\frac{\partial r}{\partial n}\big|_{\bar{s}=0} < 0$.*

As the supervisor obtains more signals, she can more accurately estimate the agent's true performance, leading to greater deviations from her prior expectation m. That is, with a larger number of signals n, signals $\bar{s}$ below the average performance m result in lower assessments, while signals above average performance yield higher assessments. Consequently, rating differentiation (the slope of the rating function) increases.

The model also predicts that – while rating differentiation increases – average ratings should remain unchanged when there are more signals, that is, $E\left[\frac{\partial r}{\partial n}\right] = 0$. In contrast, in the closely related model by Golman and Bhatia (2012, GB), a higher precision in the supervisor's signal (which here corresponds to an increase in the number of observed signals) leads to less leniency, i.e. $E\left[\frac{\partial r}{\partial n}\right] < 0$. The main difference is that GB assume an asymmetry in the (psychological) cost of giving an inaccurate rating: Supervisors have a larger cost when giving a rating below the true performance than when giving a rating above the true performance.[13] With less precise signals, errors are more likely and hence the supervisor shifts ratings upwards. When signals become more precise, this upwards

[12] See Ockenfels et al. (2020) for a similar analysis in the context of a multirater setting.

[13] To be precise, the key difference between their model and framework analyzed here (besides using a linear instead of a quadratic functional form) is that they assume an asymmetry in the supervisor's disutility from deviations which in their framework is given by

$$\begin{cases} -\lambda(a-r) & \text{if } r < a \\ -(r-a) & \text{otherwise,} \end{cases}$$

with $\lambda > 1$.

bias is reduced. Thus, we would expect a negative overall effect of higher signal precision on the ratings in the GB model and no effect in the PT framework applied here.

The model also makes a prediction regarding the role of social preferences:

Hypothesis 4: *Rating leniency is higher when supervisors have stronger social preferences η. This effect is stronger when ratings determine bonus payments ($\beta > 0$).*

The intuition is straightforward: when a supervisor cares more for the well-being of the agent, she prefers to assign higher ratings. This tendency is particularly pronounced when bonus payments are in place, as higher ratings then translate into a material benefit for the assessed agent.

Next, we examine how these parameters affect the *rating error* and the *informativeness of the ratings*. The *rating error,* measured as the the squared deviation between the rating and true performance, gauges the usefulness of the ratings in assessing performance when taken at face value. However, this aspect should not be confounded with the informativeness of ratings, which pertains to their usefulness in decision-making processes. Even biased ratings can be informative when the bias is predictable (e.g., when ratings exceed true performance by a known constant, performance can be perfectly inferred by subtracting this constant). However, if biases cannot be predicted from observable data, the ratings become less informative. We use two metrics to assess rating informativeness. First, we consider the coefficient of determination (R^2), a standard measure for the quality of predictions. Specifically, we determine the share of the overall variance in true performance that can be predicted using the ratings. As already stressed by Prendergast and Topel (1996), rating biases induced by performance pay can be harmful when they distort personnel decisions. Thus, as a second measure of rating informativeness, we consider the expected profits of a rational employer who makes promotion decisions based on the information in the ratings. For this, consider an employer who observes the rating and then either retains the agent in job 1 where output is equal to a, or promotes the agent to job 2 where output is $2a - m$, i.e. where marginal returns to performance are larger but low performance can be more detrimental.[14] The firm optimally promotes the agent whenever $\hat{a} = E[a|r] > m$ and expected profits are equal to[15] $E[a] + \Pr(\hat{a} > m) E[a - m | \hat{a} - m > 0]$.

We can show:

Proposition 2. *(i) The expected squared rating error is*

$$E\left[(a-r)^2\right] = \frac{\sigma_a^2 \sigma_\varepsilon^2}{n\sigma_a^2 + \sigma_\varepsilon^2} + \frac{(\beta+b)^2}{(\gamma+\lambda)^2}\left(\sigma_\eta^2 + m_\eta^2\right). \tag{3}$$

[14]The task assignment technology is similar to the one applied in Prendergast and Topel (1996), where profits are equal to a in job 1 and $-a$ in job 2 and which leads to qualitatively the same comparative statics results, but the technology we use avoids the issue that higher performance is detrimental in one task, and is easier to map to an optimal promotions framework such as Gibbons and Waldman (1999).

[15]Note that this is equivalent to $\Pr(\hat{a} \leq m) E[a | \hat{a} \leq m] + \Pr(\hat{a} > m) E[2a - m | \hat{a} > m]$.

(ii) The coefficient of determination when predicting true performance based on ratings is

$$R^2_{a|r} = 1 - \frac{V[a|r]}{V[a]} = \frac{1}{1 + \left(\frac{\beta+b}{\gamma+\lambda}\right)^2 \sigma^2_\eta \left(\frac{1}{\sigma_a} + \frac{\sigma^2_\varepsilon}{n\sigma^3_a}\right)^2 + \frac{\sigma^2_\varepsilon}{n\sigma^2_a}}. \tag{4}$$

(iii) The expected profits from optimal job assignment are

$$m + \frac{1}{\sqrt{2\pi}} \sqrt{\sigma^2_a R^2_{a|r}}. \tag{5}$$

Proof: See Appendix.

Importantly, all three metrics lead to analogous comparative statics predictions:

Hypothesis 5: *Rating quality (as measured by smaller rating errors, higher predictive power of ratings and higher profits from optimal job assignments) is larger when (i) agents receive no bonus payments, (ii) supervisors are rewarded for accuracy and (iii) when they observe more performance signals. (iv) Performance pay decreases rating quality to a smaller extent when supervisors are rewarded for accuracy.*

As Prendergast and Topel (1996) have already shown, rating errors increase when agents receive bonus payments for two reasons: First, as we have seen before, bonus payments induce rating leniency, leading to an upward bias. Second, when supervisors' social preferences vary (i.e. $\sigma^2_\eta > 0$) and are imperfectly known, additional noise is introduced to the rating as a performance signal. While the former effect can be corrected due to its predictable bias (when m_η is known), the latter cannot. This inability to adjust ratings when supervisor social preferences are imperfectly known undermines the informativeness of ratings.

3 Experimental Design

Our experiment consists of three parts, with data collection for each part completed before the next begins. In Part 1, subjects (called workers) perform a real effort task. After Part 1 is completed, another set of subjects (called supervisors) receives noisy information about worker performance and submits a rating about a worker's performance. After Part 2 is completed, the workers from Part 1 are informed of their ratings. We first describe the tasks in the three parts. The payoffs as well the determination of the supervisor's signal will be detailed in the subsequent treatment descriptions.[16]

[16]Screenshots of the experiment, including instructions, comprehension questions and decision screens, can be found in the Online Supplementary Material.

3.1 Tasks

Part 1

Workers perform a real-effort task. This Entry Task consists of entering text contained in hard-to-read images, similar to "captchas". Workers see 10 consecutive pages with 10 images on each page. Each page has one of five time limits: 17, 19, 21, 23, or 25 seconds.[17] There is one practice page without time limit, such that workers can familiarize themselves with the Entry Task. Performance on the practice page is not relevant for the assessment in Part 2.[18] There are no treatments in Part 1. The subjects are informed in the instructions that their work will be rated by other MTurk worker(s) and that their payment may depend on the rating.[19] Workers also fill in a demographics questionnaire.

Part 2

The second part of the experiment is performed by a different set of subjects in the role of supervisors whose main task is to rate a worker from Part 1. One worker from Part 1 is randomly and anonymously matched to each supervisor in Part 2. At the beginning of Part 2, supervisors see the practice page and work on two pages of the Entry Task.[20]

Supervisors learn about the respective treatment (as explained in detail below) and then perform the Rating Task. They receive noisy information about the performance of their matched worker and are asked to rate the worker using an integer rating $r \in [0\%, 100\%]$ (cf. Figure 1). To perform the rating, supervisors are shown the number of correctly entered images from a randomly drawn subset of the 10 pages completed by the worker (the size of this subset depends on the treatment). This information is displayed in a table, with rows for each page of the Entry Task. For pages in the signal subset, the table shows the number of correctly entered images, while no information is provided for pages outside the subset. Supervisors also see a histogram of all workers' average performances along with the mean and standard deviation. They are informed that the rating should reflect the worker's performance across all 10 pages of the Entry Task, and that performance refers to the percentage of correctly entered images.

[17] Each of these time limits occurs exactly twice in randomized order. The order of the time limits over the Entry Task's 10 pages is the same for all subjects. The time limit for the upcoming page is announced during a 5-second countdown before the page starts. The purpose of the time limit is to make information about the number of correctly entered images on a randomly selected page less informative about the worker's overall performance.

[18] After the Entry Task is finished, we elicit the workers' beliefs about their performance on all 10 pages.

[19] Note that this was necessary to avoid deceiving subjects about their payoff function, as treatments were only assigned after Part 1 was finished. Supervisors, however, learned the actual payoff functions of workers and themselves at the beginning of the experiment. The treatments are described in more detail below.

[20] One of the example pages has the shortest time limit of 17 seconds while the other page has the longest time limit of 25 seconds. The purpose is to ensure that supervisors have a good understanding of the real-effort task. We also elicit supervisors' beliefs about their own performance on the two example pages.

After the Rating Task, we assess the supervisors' social preferences towards the worker population using the incentivized Social Value Orientation slider task (SVO, Murphy et al., 2011). This task involves a randomly chosen worker who completed Part 1 (excluding the one they rated) as the matched partner. It comprises multiple dictator games where a participant decides on a monetary allocation between themselves and a matched partner, with an increase in the partner's payoff potentially reducing the participant's own payoff.[21] In terms of a utility function, SVO represents the weight a participant places on another participant's payoff relative to their own, linking it to η in our model. This social preferences parameter indicates the extent to which the supervisor values the worker's utility relative to their own. Generally, a larger SVO suggests either inequality aversion, a preference for maximizing joint surplus, or altruism. In our experimental design, where the supervisor's payoff exceeds the worker's,[22] all motives favor increasing the other's payoff. Hence, we propose that larger SVO corresponds to a larger value of η.

We also ask supervisors to write an explanation on how they determined the rating.[23]

Part 3

Workers who completed Part 1 and received a rating in Part 2 are invited per email to participate in Part 3. In this part, workers learn their ratings, their actual performance and their payment.[24]

3.2 Treatments

Our treatments are implemented in Part 2. Given the hypotheses derived in section 2, we vary whether workers are paid according to the rating or not (P and NP), whether supervisors are paid according to their rating accuracy or not (A and NA), and whether supervisors observe a subset of 1 or 4 pages out of the 10 pages the workers have worked on (S1 and S4). We employ a 2x2 design for performance pay and accuracy pay with low signal precision (S1). For high signal precision (S4), workers are always paid according to the rating and we only vary accuracy pay. Altogether, we conduct 6 treatments (see Table 1). We randomly assign workers (and hence matched supervisors) to the six treatments after they completed Part 1. Treatment assignment was stratified to obtain similar performance distributions across treatments.

[21] We utilize the six primary items from Murphy et al. (2011), with each point valued at \$0.01.

[22] Given a very large rating error, the supervisor could earn less than the matched worker in the treatments with accuracy pay. However, in our study, 99.4% of supervisors earn more than their matched worker.

[23] Supervisors also fill in a short reciprocity questionnaire (Dohmen et al., 2009), the Big Five Inventory (Rammstedt and John, 2005), and the same demographics questionnaire as the workers in Part 1.

[24] We also elicited their satisfaction with the rating and they completed the incentivized SVO slider measure with their supervisor as the recipient to measure their social preferences towards their supervisor as a reaction to their supervisor's rating and the associated payment (see Section A.2 in the Appendix for an analysis of this data).

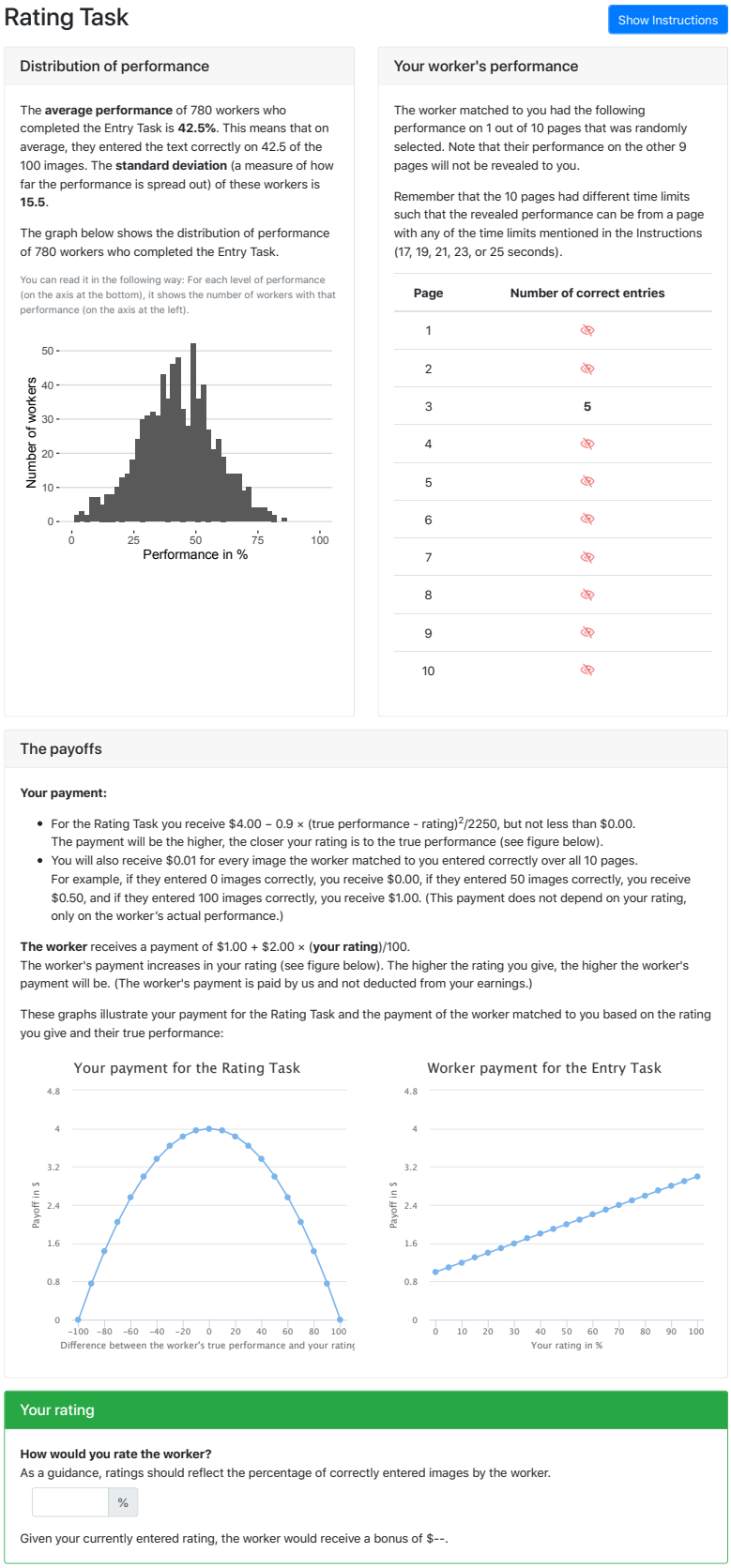


Figure 1: Screenshot of the Rating Task (treatment P-A-S1)

Table 1: Treatments

Name	λ	β	n
NP-NA-S1	0	\$0	1
NP-A-S1	0.0004	\$0	1
P-NA-S1	0	\$2	1
P-A-S1	0.0004	\$2	1
P-NA-S4	0	\$2	4
P-A-S4	0.0004	\$2	4

In the treatments where supervisors are paid according to their rating accuracy (the A-treatments), they receive a monetary payoff of $\$4 - 0.0004(r-a)^2 + \$0.01a$, which depends on the squared rating error (that is, the squared difference between true performance a and rating r). In the treatments without accuracy incentives (the NA-treatments), their payment is equal to $\$4 + \$0.01a$. In all treatments, supervisors receive \$0.01 for each image their matched worker entered correctly.[25] Workers receive a fixed payment of \$1 in all treatments. In the treatments where they are paid according to the rating (the P-treatments), they additionally earn a bonus of $\frac{r}{100} \cdot \$2$.

Supervisors are informed about both their own and the worker's payoff function as determined by the respective treatment. In case the supervisor's payoff depends on accuracy, supervisors are shown a plot of their payoff as a function of their squared rating error. In case the worker's payoff depends on the rating, supervisors are shown a plot of the worker's payoff as a function of their rating (both plots are shown in Figure 1). In Part 3, workers only learn their own payoff function, i.e. whether their own performance depends on the rating or not. They are not aware of how many signals the supervisor saw or whether the supervisor's payment depends on accuracy.

3.3 Procedures

We conducted the experiment online on Amazon MTurk, a website for crowdsourced labor.[26] On MTurk, so-called requesters announce a task or a study (called HIT, human intelligence task) using a short title, a description, and a reward for completing the task.[27]

[25] We include this payment to ensure supervisors, like in real-world relationships, benefit from and care about the workers' performance. Note also that supervisors do not bear the cost of the payment to the worker, reflecting that in most field settings, supervisors are themselves employees who carry out the rating task, but the payment itself is made by a firm as their employer.

[26] See e.g. Arechar et al. (2018); Horton et al. (2011); Paolacci et al. (2010) for running experiments on Amazon MTurk.

[27] Our study was advertised with the title "Academic study (~X minutes, additional bonus)", where X was the duration we estimated for each part (see below), and the description "Participate in an academic study on human decision-making and earn money. Read the preview for further information.". The preview page is described in greater detail below. The reward was \$0.50 in all three parts.

The reward is the same for all participants for a given HIT and is comparable to the show-up fee paid in a physical laboratory. Individual payments that depend on decisions during the experiment can be made in the form of a so-called bonus. We restricted participation to MTurk workers who have completed at least 1000 HITs on MTurk and who have an approval rate of at least 98% in order to ensure high data quality. The experiment was computerized using oTree (Chen et al., 2016) and embedded in the MTurk website.

The preview page of the experiment contained a short description of the study, the estimated duration,[28] information about the possible compensation, and contact details of the authors conducting the study. In parts 1 and 2, this page also contained the technical requirements[29] and an informed consent form which the subjects needed to accept in order to start the experiment. They were made aware that after reading the instructions, they have to answer comprehension questions to ensure their understanding of the instructions, and that if they do not answer a question correctly after the third attempt, they will be excluded from further participation and payment. Subjects could review the instructions during answering the comprehension questions and throughout the rest of the experiment. Subjects in Part 1 were also informed that they will only receive their bonus if they also participate in the third part of the study within 4 weeks of receiving the invitation email.

Part 1 was online from 2019/11/18 to 2019/11/20, Part 2 was run from 2019/11/21 to 2019/11/22, and Part 3 was available from 2019/11/24 until 2019/12/31. Parts 1 and 2 were accessible on MTurk between 8am Eastern Time until 8pm Pacific Time in order to minimize variations in demographic composition over time of day (see Casey et al., 2017). Subjects could only participate either both in parts 1 and 3 (and in 3 only if they had completed Part 1 first) or in Part 2. Subjects could not participate more than once, ensured by using MTurk qualifications. Once they accepted the task, subjects had 60 minutes to complete the experiment in parts 1 and 2. In case they exceeded the time limit, they were excluded from participation, did not receive a payment and their slot was made available to another MTurk worker. There was no such time limit in Part 3.[30] On average, Part 1 lasted 11 minutes, Part 2 took 16 minutes to complete, and subjects spent 4 minutes in Part 3. The average payment to workers (supervisors) was \$4.16 (\$6.32), yielding hourly wages of \$16.64 (\$23.70) well above US minimum wage standards. We kept Part 1 online until we had 780 participants, translating into 130 worker/supervisor groups per treatment.[31]

[28] Our estimates were 13, 13, and 4 minutes for parts 1, 2, and 3, respectively.

[29] Subjects needed a screen of at least 13", a physical keyboard, and a browser with JavaScript to ensure a level playing field for the Entry Task.

[30] Using a time limit is necessary on MTurk, as it is common for workers to accept tasks without starting to work on them, blocking slots for workers who are directly available and delaying the experiment.

[31] Due to technical difficulties in Part 2, 8 supervisors rated a worker who already had received a rating by a different supervisor. For these 8 workers, we randomly picked one of the two supervisors for use in Part 3. The unused supervisor was paid according to his/her decisions but was not the recipient of the worker's SVO decision in Part 3.

Table 2: Mean and variance of rating and performance

Treatments	Performance		Rating	
	Mean	Variance	Mean	Variance
NP-NA-S1	42.55	242.56	43.22	638.29
P-NA-S1	42.79	242.69	51.49	633.26
P-A-S1	42.44	242.51	46.66	551.06
NP-A-S1	42.37	241.27	42.58	426.60
P-NA-S4	42.66	240.61	50.42	561.55
P-A-S4	42.25	242.05	45.55	445.34

After we gathered a rating for each worker in Part 2, we emailed invitations to the workers from Part 1 to participate in the final part. In Part 3, 764 of 780 subjects from Part 1 returned. The attrition rate of 2.1% did not differ between the treatments.

As to the demographics of our sample, 50.5% of our subjects are female. Average age was 37.9 years, with a minimum (maximum) of 19 (76) years. 16.7% spend less than 5 hours per week on MTurk, 35.3% between 5 and 10 hours, and 48% spent more than 10 hours per week on MTurk.

4 Results

4.1 Descriptives

We begin our analysis by reporting descriptive statistics on the agents' performance and the assigned performance ratings. Table 2 reports means and variances for our six treatments. The random assignment of supervisors to the treatments indeed generated very similar distributions of the underlying performance outcomes. However, the distribution of ratings varies strongly between the treatments. We explore these treatment differences and their drivers in detail in the following sections.

4.2 Performance Pay and Rating Leniency

As key benchmark case, we first consider average ratings in the baseline condition NP-NA-S1, where there is no performance pay and no reward for accuracy. A first important observation is that we do not observe sizable rating leniency in that case, as the average rating of 43.22 is only slightly and insignificantly larger than the average performance of 42.55 ($p = 0.7508$, two-sided t-test). Recall that the formal model predicts rating leniency in this case only if supervisors internalize a potential non-monetary psychological benefit agents receive from a higher ratings (i.e. $b > 0$). The fact that we observe little leniency when there is no performance pay indicates that those benefits – if anything – only play

a minor role.

We now investigate the effect of agent's performance pay on the supervisors' rating behavior in comparison to this benchmark setting, testing Hypothesis 1. To accomplish this, we compare treatment NP-NA-S1 with treatment P-NA-S1, in which agents receive a bonus based on the supervisors' evaluation. By equation (2), the model predicts that the rating should be higher when performance pay is in place ($\frac{\partial r}{\partial \beta} > 0$), but the slope of the rating function with respect to the signal should be unaffected ($\frac{\partial^2 r}{\partial \tilde{s} \partial \beta} = 0$). Table 3 reports regressions of the rating on a dummy for the use of performance pay for the agent (column (1)), as well as the signal observed by the supervisor (column (2)), and an interaction term between both (column (3)). The experimental results are well in line with Hypothesis 1. Supervisors become substantially more lenient when there is performance pay. Their ratings increase by 8.3 percentage points (or by about 20%) when performance pay for the agent is in place. This effect persists when we control for the realization of the signal in column (2). In line with the model, we find no evidence of an effect of performance pay on rating differentiation: The interaction term of signal and performance pay in column (3) is small and not significantly different from zero.

Panel (a) of Figure 2 shows regression lines and 95% confidence intervals for the relation between signal and rating in the treatments without accuracy pay. It also depicts the optimal rating function for supervisors without social preferences ($\eta = 0$) as a dashed black line. While the ratings in the treatment without exogenous incentives NP-NA-S1 are close to the optimal rating function, the introduction of performance pay for the agents shifts the ratings upwards. However, performance pay does not affect the slope. Taken together, we find evidence for Hypothesis 1 as the introduction to performance pay leads to more lenient ratings.

4.3 Supervisors' Incentives for Accuracy and Agents' Performance Pay

In a next step, we study the interplay between the use of performance pay for the agent and the provision of incentives for accuracy, testing Hypothesis 2. The model implies that providing incentives for accuracy should reduce ratings ($\frac{\partial r}{\partial \lambda} < 0$) and that this effect should be stronger when performance pay is in place ($\frac{\partial^2 r}{\partial \lambda \partial \beta} < 0$). However, it is important to note that a reward for accuracy only should reduce leniency when there is leniency in the first place. The model predicts that – when there is no performance pay – this occurs only when supervisors internalize a potential non-monetary psychological benefit b of higher ratings for the agent. As argued in the previous subsection, there is little evidence for this effect of b. Hence, we should expect that a reward for accuracy reduces leniency in particular when there is performance pay.

We start our analysis by regressing the rating on the signal and a treatment dummy for accuracy payment, separately for the treatments without and with performance pay

Table 3: The effect of performance pay (no incentives for accuracy)

	(1) Rating	(2) Rating	(3) Rating
Performance pay	8.277***	9.732***	12.74**
	(3.127)	(2.672)	(6.268)
Signal		0.573***	0.603***
		(0.0653)	(0.0863)
Signal × Performance pay			-0.0717
			(0.132)
Constant	43.22***	18.30***	17.00***
	(2.216)	(3.096)	(3.778)
Observations	260	260	260
R^2	0.026	0.274	0.275

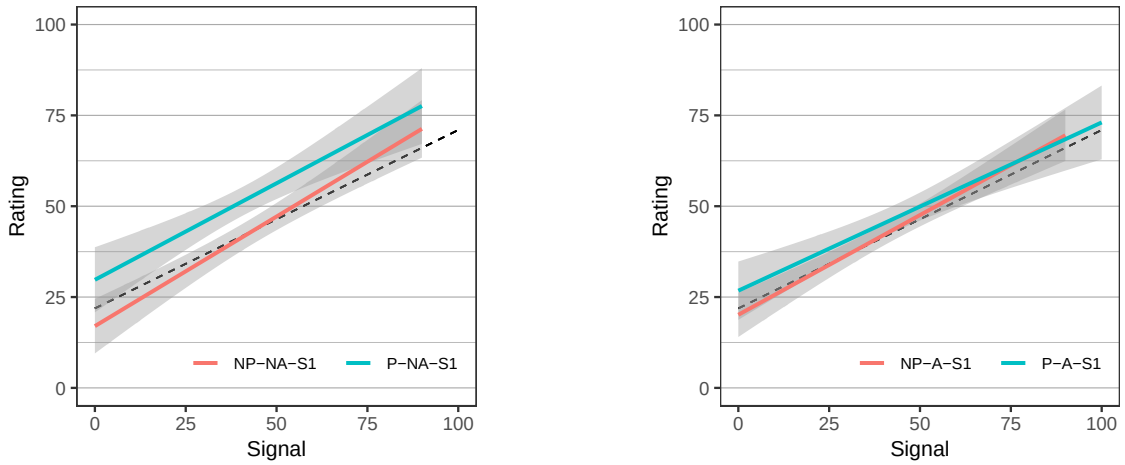
Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable is the assigned rating. *Performance pay* is a dummy indicating whether the rating determines a bonus paid to the agent. *Signal* is the value of the signal observed by the supervisor. Data from treatments NP-NA-S1 and P-NA-S1.

(columns (1) and (2) of Table 4). The effect of accuracy pay is only significant in the treatments with performance pay, where it reduces the average rating.

Column (3) combines the data from the two previous models and includes an interaction effect between accuracy and performance pay. Again, it shows that the leniency introduced by performance pay is indeed reduced when supervisors' pay depends on their rating error. About 2/3 of the leniency effect disappears in this case. These effects can also be seen in panel (b) of Figure 2: If there is accuracy pay for the supervisor, then performance pay for the agent does not increase ratings substantially (corresponding to an average marginal effect of performance pay in the model in column (3) of 2.96 ($p = 0.213$) when accuracy pay is in place). A comparison of the regression lines for P-NA-S1 and P-A-S1 across panels (both in blue) illustrates the downward shift caused by the introduction of accuracy pay when performance pay is in place. Taken together, we do not find support for the strong formulation of Hypothesis 2, which states that accuracy pay always reduces ratings. However, we do find support for the second part of the hypothesis that suggests that accuracy pay reduces ratings when agents' payments are linked to these ratings. The two insignificant interaction terms in column (4) are also in line with the model, as they indicate that there is no evidence of accuracy pay or performance pay affecting the slope of the rating function (i.e. $\frac{\partial r}{\partial s}$, or the extent to which the rating depends on the signal).

4.4 More Signals

In a next step, we include the treatments where we varied the precision of the signal observed by the supervisor, testing Hypothesis 3. The model implies a shift in both the



(a) Treatments without accuracy pay

(b) Treatments with accuracy pay

Figure 2: The effect of performance pay on ratings in treatments with one signal. The dashed black line denotes the optimal rating function for supervisors without social preferences ($\eta = 0$).

intercept ($\frac{\partial^2 r}{\partial \bar{s} \partial n} > 0$) as well as the slope of the rating function ($\frac{\partial r}{\partial n}\big|_{\bar{s}=0} < 0$) such that rating differentiation increases. That is, the ratings should vary to a stronger extent with the observed signals. Put differently, the intercept of the optimal rating function is decreasing in n and its slope is increasing in n. Hence, signals below average performance m should lead to lower ratings in the S4 treatments than in the S1 treatments, while above-average signals should lead to higher ratings.

Table 5 shows the results of regressions of the rating on the signal average interacted with a dummy for the treatment with four signals, allowing for a different slope and a different intercept in treatments with higher signal precision. The first two columns show the results separately for the treatments without and with accuracy incentives, while column (3) uses pooled data. In all three columns, we see that when supervisors have four signals instead of one at their disposal, the intercept is decreased by about 10 percentage points. At the same time, the slope of the rating function becomes more steep, as seen in the interaction effect of Four signals and the signal average. Hence, rating differentiation increases. The effects are significant in the pooled data and in the treatments with accuracy incentives. The point estimates are of very similar magnitude in the treatments without accuracy incentives but insignificant, as standard errors are larger.

These results thus support Hypothesis 3 and are in line with similar experimental findings in Ockenfels et al. (2020). We also present these results graphically in Figure 3, which mirrors Figure 2 and adds the optimal rating function for supervisors without social preferences (η =0) with four signals as a black dotted line.

Table 4: The interaction between performance pay and incentives for accuracy

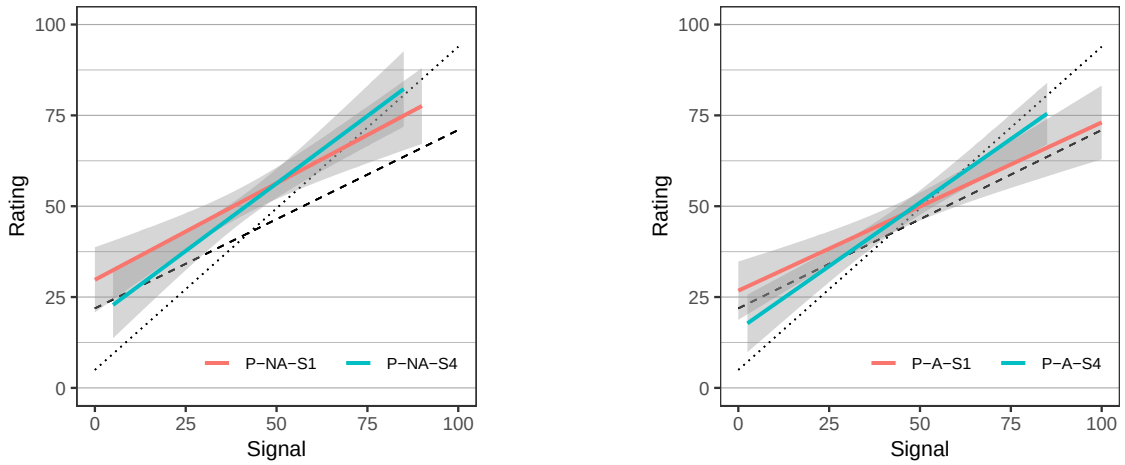
	(1) No perf. inc.	(2) Perf. inc.	(3) Pooled	(4) Pooled
Signal	0.578***	0.494***	0.540***	0.606***
	(0.0541)	(0.0689)	(0.0431)	(0.0750)
Accuracy pay	0.837	-5.857**	0.739	3.395
	(2.325)	(2.723)	(2.332)	(4.084)
Performance pay			9.647***	13.04***
			(2.688)	(4.576)
Performance pay × Accuracy pay			-6.690*	-6.708*
			(3.581)	(3.578)
Signal × Accuracy pay				-0.0608
				(0.0850)
Signal × Performance pay				-0.0789
				(0.0864)
Constant	18.09***	31.27***	19.77***	16.87***
	(2.681)	(3.722)	(2.441)	(3.392)
Observations	260	260	520	520
R^2	0.335	0.194	0.271	0.274

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable is the assigned rating. *Performance pay* is a dummy indicating whether the rating determines a bonus paid to the agent. *Accuracy pay* is a dummy indicating whether the supervisor is rewarded for accuracy. *Signal* is the value of the signal observed by the supervisor. Data in (1) from treatments NP-NA-S1 and NP-A-S1, in (2) from treatments P-NA-S1 and P-A-S1, and in (3) and (4) from treatments NP-NA-S1, NP-A-S1, P-NA-S1 and P-A-S1.

Table 5: The effect of signal precision (treatments with performance pay)

	(1) No acc. inc.	(2) Acc. inc.	(3) Pooled
Signal average	0.532***	0.463***	0.494***
	(0.100)	(0.0945)	(0.0688)
Four signals	-10.62	-10.74*	-10.76**
	(7.110)	(6.192)	(4.693)
Four signals × Signal average	0.211	0.237*	0.227**
	(0.148)	(0.127)	(0.0969)
Accuracy pay			-5.394***
			(1.806)
Constant	29.74***	26.75***	31.06***
	(5.002)	(4.670)	(3.560)
Observations	260	260	520
R^2	0.213	0.243	0.234

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable is the assigned rating. *Signal average* is the average value of the signals observed by the supervisor. *Four signals* is a dummy indicating that the supervisor observes four rather than one signal. *Accuracy pay* is a dummy indicating whether the supervisor is rewarded for accuracy. Data in (1) from treatments P-NA-S1 and P-NA-S4, in (2) from treatments P-A-S1 and P-A-S4, and in (3) from treatments P-NA-S1, P-NA-S4, P-A-S1 and P-A-S4.



(a) Treatments without accuracy pay

(b) Treatments with accuracy pay

Figure 3: The effect of signal precision on ratings. The dashed (dotted) black line denotes the optimal rating for supervisors without social preferences ($\eta = 0$) with one signal (four signals)

So far, we have investigated the effect of having access to more precise information on the shape of the supervisors' rating function (i.e., conditional on the observed signal). It is also instructive to investigate the effect of the additional information on the average rating (unconditional on the realized signal). As noted above, the framework presented in Section 2 predicts that the precision of the available information does not affect overall rating leniency. Thus, the average rating should not differ between the treatments with one and four signals. Golman and Bhatia's (2012) related model, however, predicts such a difference. When, as is assumed in that model, supervisors' disutility from inaccurate ratings is asymmetrically affected by deviations above and below the true performance, and supervisors have a stronger urge to avoid negative deviations,[32] more precise signals should also reduce rating leniency.[33]

However, in our data we find no significant effect of the signal precision on the average rating: When there are accuracy incentives, the average rating in S4 (S1) is 45.5 (46.7), and without accuracy incentives the average rating in S4 (S1) is 50.4 (51.5). While

[32]There are some lab experiments comparing leniency and severity errors (i.e. errors that either lead to an upward or downward bias in evaluations). For instance, Dickson et al. (2009) find that in in public goods games with punishment and noisy information about contributions a monitoring technology that makes severity errors decreases contributions more than one that makes leniency errors. Moreover, subjects have a larger willingness to pay to play in an environment that makes leniency errors compared to one with severity errors (Markussen et al., 2016). In a related principal-agent experiment, Marchegiani et al. (2016) show that a monitoring technology that creates leniency errors decreases effort by an agent less than one with severity errors.

[33]There is indeed evidence from previous laboratory experiments in the related setting of feedback on online markets showing that uncertainty about the seller's intentions when receiving subpar quality leads to more lenient ratings (Rice, 2012; Bolton et al., 2019).

the average ratings are larger in the treatments with a less precise signal as predicted by Golman and Bhatia's (2012) model, the differences of about −1.2 percentage points are not significant and relatively small (cf. the regression results in Table A.1 in the Appendix). By comparison, the introduction of performance pay for the worker increases average ratings by 8.3 percentage points (see Table 3). Thus, we find no evidence of uncertainty increasing average ratings in our experimental setting.

4.5 Supervisors' Social Preferences

According to Hypothesis 4, stronger social preferences η lead to an increase in average ratings as $\frac{\partial r}{\partial \eta} > 0$. This effect should be more pronounced when the rating determines a performance bonus as $\frac{\partial^2 r}{\partial \eta \partial \beta} > 0$. Moreover, the model also predicts that both effects should be less pronounced when the supervisor receives a bonus for accurate ratings λ as in this case, the supervisor's ratings are predicted to vary less with her social preferences.

We make use of the exogenous variation of the supervisor assignment to test whether there is a correlation between the supervisor's social preferences and the assigned ratings.[34] To measure the supervisors' social preferences towards the worker population (or η in the language of our model), we elicited their Social Value Orientation. Throughout our analysis, we use the standardized SVO angle, which represents the weight of the receiver's payoff compared to one's own in the utility function (for details on the derivation of the SVO angle, see Murphy et al., 2011).[35]

Table 6 shows regressions of the rating on the supervisor's SVO interacted with a dummy for the use of performance pay, controlling for the signal average. Column (1) considers the treatments without incentives for accuracy, where the effects of performance pay should be strongest (NP-NA-S1 and P-NA-S1). In line with our hypothesis, we find that supervisors with higher social preferences indeed provide significantly more lenient ratings when there is no performance pay. With performance pay, our model predicts a stronger effect of social preferences (i.e. that $\frac{\partial r}{\partial \eta \partial \beta} > 0$). However, the interaction effect between SVO and the performance pay dummy is not significantly different from zero.[36] Hence, while we find that performance pay triggers rating leniency, we find no evidence that this effect is stronger for more prosocial supervisors.

Column (2) shows the results for the treatments where the supervisor is rewarded for accuracy (NP-A-S1 and P-A-S1). In line with the hypothesis that prosocial preferences

[34] Kane et al. (1995) find evidence in line with the conjecture that a supervisor's tendency to provide lenient ratings tends to be driven by stable personality traits. Breuer et al. (2013) find that supervisors tend to assign better ratings at the same level of objective performance to workers with whom they have worked for a longer time before.

[35] The SVO consists of supervisor's choices in a series of dictator games. We discuss the SVO elicitation and its relation to η in Section 3.1 on the experimental design.

[36] The coefficient even exhibits a negative sign, but the respective standard error is large.

Table 6: The association between supervisors' social preferences and ratings

	(1) No acc. inc.	(2) Acc. inc.
SVO (std.)	5.612***	0.209
	(1.851)	(1.759)
Performance pay	9.659***	3.121
	(2.644)	(2.402)
SVO (std.) × Performance pay	-3.673	-1.504
	(2.637)	(2.532)
Signal average	0.568***	0.506***
	(0.063)	(0.056)
Constant	18.544***	21.870***
	(3.117)	(2.711)
Observations	260	260
R^2	0.301	0.266

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable is the assigned rating. *SVO (std.)* is the supervisor's standardized SVO angle. *Performance pay* is a dummy indicating whether the rating determines a bonus paid to the agent. *Signal average* is the average value of the signals observed by the supervisor. Data from treatments with 1 signal (S1).

should affect ratings less under accuracy incentives, the coefficient of SVO is much weaker. In fact, it even does not have any sizable predictive power for rating leniency in this case, neither with nor without performance pay.

Therefore, a key observation of this analysis is that, while performance pay leads to more leniency when there are no incentives for accuracy, this increase is not mainly driven by more prosocial supervisors.[37] This indicates that our ex-ante interpretation of the model, according to which prosocial preferences only matter through the weight the supervisor gives to the payoffs of the evaluated workers, has been too narrow. Indeed, it appears likely that prosociality has richer consequences. We thus explore other dimensions in how more prosocial supervisors differ in their rating behavior from less prosocial ones.

Recent research in psychology has shown that prosocial subjects (also measured by SVO) invest more time and effort to assess the consequences of their actions for others (Bieleke et al., 2020).[38] Applied to our setting, this finding suggests that prosocial super-

[37]When there is no accuracy pay for the supervisor but performance pay for the worker, supervisors can raise workers' payoffs at no material costs for themselves, which may explain why less prosocial supervisors also tend to increase their ratings in this case (assuming that their intrinsic preferences for accuracy are small). Bruhin et al. (2019), for instance, estimate heterogeneity in social preferences with a structural model and detect hardly any purely selfish individuals, but many "moderately altruistic" ones which choose prosocial actions only when the costs are small.

[38]Similarly, Grosch and Rau (2017) show experimentally that prosocial subjects are more honest.

Table 7: The association between rating diligence and supervisors' social preferences

	(1) Duration	(2) Explanation length
SVO (std.)	12.75***	16.28***
	(2.280)	(3.673)
Performance pay	5.334	-2.340
	(6.920)	(13.35)
Accuracy pay	15.43*	2.828
	(7.875)	(14.43)
Performance pay × Accuracy pay	-15.88	-5.135
	(9.777)	(16.94)
Four signals	12.45**	5.333
	(5.794)	(8.820)
Constant	89.41***	147.9***
	(5.222)	(10.98)
Observations	780	780
R^2	0.046	0.024

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable is the the duration of the Rating Task in seconds in (1) and the number of characters in the explanation supervisors wrote about how they determined their rating in (2). *SVO (std.)* is the supervisor's standardized SVO angle. *Performance pay* is a dummy indicating whether the rating determines a bonus paid to the agent. *Accuracy pay* is a dummy indicating whether the supervisor is rewarded for accuracy. *Four signals* is a dummy indicating that the supervisor observes four rather than one signal. Data from all treatments.

visors may also be more diligent in their evaluation behavior. To investigate whether this is indeed the case, we make use of two proxies for rating diligence. First, we recorded the time (in seconds) supervisors spent making their rating decision. Second, we included an open-ended question in the post-experimental survey asking supervisors to explain their rating decision process and counted the number of characters in each supervisor's response. To validate these proxies, we assess their correlation with the squared rating error (i.e. the squared deviation between rating and true performance, a metric we explore in more detail in subsection 4.6). Both proxies are negatively correlated with the rating error (duration: Spearman's $\rho = -0.1001$, $p = 0.0052$, explanation length: Spearman's $\rho = -0.1691$, $p \leq 0.0001$). We then regress these proxies on SVO and treatment dummies. As Table 7 shows, more prosocial supervisors indeed take more time when rating the worker (column 1) and use more characters to explain their ratings (column 2). Both findings support the view that more prosocial supervisors tend to be more diligent when performing the rating task. This suggests that social preferences influence ratings in a more complex manner than our narrow interpretation of the model suggested. It appears likely that the higher diligence also affects rating quality – a question that we

investigate in more detail in the following section.

4.6 Rating Error and the Informativeness of Ratings

In a next step, we analyze the effect of our treatments on the quality of ratings as measured by the *rating error* and the *informativeness of the ratings*. As summarized in Hypothesis 5, the model predicts that rating quality is larger when (i) agents receive no bonus payments, (ii) supervisors are rewarded for accuracy, and (iii) when supervisors observe more performance signals. Moreover, (iv) performance pay should decrease rating quality to a smaller extent when supervisors have incentives for accurate ratings.

Recall that we defined the *rating error* as the squared deviation between the rating and actual performance. To test Hypothesis 5 for this metric, we report a regression with the squared rating error as the dependent variable in column (1) of Table 8. The regression partially confirms Hypothesis 5 for this metric: When there is no reward for accuracy, performance pay increases the rating error,[39] and having access to more signals leads to significantly more accurate ratings. While we find no significant evidence that accuracy pay reduces the rating error when there is no performance pay, it does so when there is (as the sum of the Accuracy pay dummy and the interaction term Performance pay $\times$ Accuracy pay has a point estimate of -195.6, $p = 0.021$, see also the two rows at the bottom of Table 8). Taken together, we find that the the presence of a worker bonus and accuracy incentives for the supervisor need to be considered jointly: Performance pay only significantly increases rating errors when there is no accuracy pay, and accuracy pay only significantly reduces rating errors when they are sizable enough, which is the case when there is performance pay.

As laid out in section 2, rating errors are an imperfect measure of rating quality, as they measure the quality of ratings when taken at face value. In a next step, we therefore consider how the treatments affect the *informativeness* of the ratings. We proceed in two steps. First, we descriptively assess how useful the ratings are in *predicting actual performance*. That is, for each treatment, we regress each worker's actual performance on the rating they received[40] and compare the coefficient of determination (R^2) as a measure of prediction quality. Table 9 shows the R^2 values for all treatments. When there is no accuracy pay, performance pay appears to reduce the predictive power of ratings, as the respective R^2 drops from 0.180 to 0.075. This is not the case when there is accuracy pay

[39] When there is accuracy pay, however, performance pay does not significantly affect rating errors, although the coefficient is still positive and sizable (the sum of the Performance pay dummy and the Accuracy pay $\times$ Performance pay interaction term has a point estimate of 96.317, $p = 0.319$).

[40] We also looked at other machine learning algorithms such as random forests. However, simple linear regressions appear to perform better, likely because the underlying true conditional expectation function closely approximates a linear relationship. In this case, the Regression CEF Theorem (Angrist and Pischke, 2009, p. 38) applies, which shows that OLS regressions yield the best linear approximation to the conditional expectation function (in the MMSE sense).

Table 8: The effects of the treatments, supervisors' social preferences and rating diligence on rating quality

	Squared rating error			Profit		
	(1)	(2)	(3)	(4)	(5)	(6)
Performance pay	190.906*	195.481*	193.941*	-1.321*	-1.358*	-1.352*
	(103.734)	(103.495)	(103.261)	(0.712)	(0.709)	(0.709)
Accuracy pay	-101.008	-98.232	-94.245	-1.001	-1.023	-1.048
	(92.034)	(91.759)	(92.091)	(0.798)	(0.794)	(0.793)
Performance pay	-94.588	-100.299	-106.359	2.226**	2.272**	2.308**
× Accuracy pay	(124.874)	(124.076)	(122.276)	(0.973)	(0.969)	(0.962)
Four signals	-178.427**	-185.639**	-179.738**	0.950*	1.007*	0.974*
	(84.400)	(84.497)	(83.348)	(0.558)	(0.563)	(0.558)
SVO (std.)		-65.022**	-49.460		0.515**	0.432*
		(32.238)	(30.580)		(0.242)	(0.238)
Duration			-0.097			0.001
			(0.460)			(0.003)
Explanation length			-0.880***			0.004**
			(0.317)			(0.002)
Profit benchmark (centered)				0.983***	0.983***	0.984***
				(0.014)	(0.014)	(0.014)
Constant	543.208***	543.077***	681.917***	45.790***	45.791***	45.060***
	(69.237)	(68.925)	(89.227)	(0.484)	(0.481)	(0.608)
p-value of P + P × A	0.319	0.323	0.354	0.269	0.262	0.237
p-value of A + P × A	0.021	0.018	0.016	0.028	0.025	0.023
Observations	780	780	780	780	780	780
R^2	0.016	0.022	0.033	0.930	0.931	0.931

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable is the squared difference between the rating and the actual performance outcome in (1)–(3) and the employer profit measure from hypothetical job assignments based on ratings in (4)–(6). *Performance pay* is a dummy indicating whether the rating determines a bonus paid to the agent. *Accuracy pay* is a dummy indicating whether the supervisor is rewarded for accuracy. *Four signals* is a dummy indicating that the supervisor observes four rather than one signal. *SVO (std.)* is the supervisor's standardized SVO angle. *Duration* is the duration of the Rating Task in seconds. *Explanation length* is the number of characters in the explanation supervisors wrote about how they determined their rating. *Profit benchmark (centered)* is the mean-centered employer profit from hypothetical job assignments based on true performance. Data from all treatments.

(where the R^2 is even slightly larger when there is performance pay). The availability of more signals increases the R^2 both with accuracy pay (from 0.134 to 0.264) and without (from 0.075 to 0.168). We take this as evidence that prediction quality is larger when the agent receives no bonus and when there are more signals available.

Second, we use the predicted performance to compute a profit measure based on a

Table 9: Informativeness of supervisor ratings in predicting true performance across treatments

Treatment	NP-NA-S1	P-NA-S1	NP-A-S1	P-A-S1	P-NA-S4	P-A-S4
R^2	0.180	0.075	0.120	0.134	0.168	0.264

This table displays the R^2 of OLS regressions of the actual performance on the performance rating for each treatment.

hypothetical personnel decision of an employer who uses the information contained in the ratings. As described in section 2, we consider a technology where an employer can assign an agent to one of two jobs. The employer's profit depends on the agent's true performance (denoted by a) and the job assignment. When assigning a worker to job 1, her profit is equal to a, while in job 2, it is equal to $2a - m$, where m is the average performance of all agents. As shown above, the employer will optimally use the predicted performance $\hat{a}$ for job assignment, and agents predicted to perform below average will be retained in job 1, while those predicted to be above average will be promoted to job 2. Based on predicted performance $\hat{a}$ from the regressions in Table 9, we compute the profits an employer would achieve when using the ratings optimally under such a technology.

Columns (4)–(6) in Table 8 show regressions with this employer profit measure as the dependent variable. We control for the benchmark profits that an employer would achieve when making the decision based on actual performance a (which is not affected by the treatments). The results mostly are in line with the analysis for the rating error: As the regression results in column (4) show, profits from job assignments tend to be smaller when there is performance pay (and no accuracy pay) and larger when there are more signals. Accuracy pay does not significantly raise profits when there is no performance pay, but does so when performance pay is used (the point estimate of the sum of the Accuracy pay dummy and the interaction term Performance pay × Accuracy pay is 1.226, $p = 0.028$). Summing up, we find that rating quality (measured both by lower squared rating errors and larger employer profits from job assignments) is (i) reduced by introducing performance pay (without accuracy pay), (ii) increased by introducing accuracy pay (with performance pay), and (iii) increased when there are more performance signals.

Our final question is how the supervisors' social preferences affect rating quality. When sticking to the narrow interpretation of the model, higher social preferences lead to higher rating leniency, which should in turn increase rating errors.[41] But as the analysis in the previous section has shown, supervisors with stronger social preferences are also

[41] As to rating informativeness, the model's predictions are less straightforward, as even biased ratings can be informative when this "bias" is predictable. To see that, note that according to Proposition 2, a higher expected value of the social preference parameter m_η increases rating error (claim i) but not rating informativeness, as the expressions in claim ii) and iii) are unaffected by m_η. However, larger supervisor heterogeneity (as measured by larger σ_η) reduces rating quality with respect to all three claims in Proposition 2.

more diligent in their rating behavior. Columns (2) and (5) of Table 8 study this question by including SVO as an independent variable. The regression results show that rating errors are significantly smaller and rating informativeness (as measured by job assignment profits) is significantly higher when the supervisor is more prosocial. Hence, in contrast to our ex-ante expectation, social preferences do not reduce, but rather increase rating quality.

While cautioning that our experiment was not designed ex-ante to study the mechanisms on why more prosocial supervisors provide a larger rating quality, we provide exploratory evidence on this question. To perform a simple descriptive mediation analysis, we include two variables to measure rating diligence introduced in the previous section on social preferences in columns (3) and (6) of Table 8: The time supervisors spent on the Rating Task and the length of the explanation on how they determined their rating. While the coefficient of rating duration is not significantly different from zero, longer explanations are significantly positively associated with rating quality as measured both by a lower rating error and higher hypothetical profits from job assignments. The inclusion of these variables in the regression models also reduces the coefficients of supervisor SVO both in size and level of statistical significance. This can be seen as an indication that the SVO effect is indeed (at least partly) driven by rating diligence.Conclusion

We have shown that a standard formal framework to model subjective performance evaluations by a rational decision-maker can organize our experimental results quite well in several dimensions. Performance evaluations become more lenient and less informative when supervisors determine bonus payments. However, rewards for accuracy counteract this effect: They reduce leniency and increase the informativeness of ratings in this case. Moreover, in line with rational Bayesian updating, we find that when more information is available, supervisors follow their observed performance signals to a stronger extent, which leads to less compressed evaluations and a higher informativeness of ratings.

Our results have several implications for the practice of performance management. For one, the results further support the claim that there is a tension between different evaluation purposes, for instance, that evaluations that are used to assign bonuses are less useful for personnel decisions. Moreover, particularly in this case, providing incentives for accurate evaluations is crucial. A further key implication of our results concerns the question whether prosocial preferences of the supervisor undermine or foster the informativeness of ratings. While our ex-ante expectation was that prosocial preferences lead to more lenient and, in turn, less informative ratings, we have actually found a more complex relationship. Indeed, prosociality is associated with more leniency on average, but supervisors who have stronger prosocial preferences are also more diligent and appear to invest more effort in the rating task. In fact, as our results have shown, the latter effect outweighs the former: more prosocial supervisors even provide ratings that

have smaller errors and are more informative. Hence, rather than fearing that prosocial managers distort personnel decisions due to providing merely generously inflated ratings, firms may instead expect them to provide more reliable information.[42]

In our experimental design, the degree of subjectivity is relatively low, as supervisors were asked to estimate the true performance based on a noisy signal. In a real-world setting, the supervisors' evaluations could be less constrained and thus more influenced by personal bias or discretion. Consequently, it may be argued that the effects of our treatments on rating leniency are likely to be more conservative estimates in this respect.

There are several avenues to explore in future research on the role of social preferences for subjective performance evaluations. Our finding that more prosocial supervisors are more diligent in their rating behavior appears to be well aligned with a growing literature in behavioral economics on preferences for meritocratic fairness. According to meritocratic fairness norms, individuals who exert higher effort also deserve higher payoffs (compare for instance Cappelen et al., 2007, Cappelen et al., 2022). In a recent experiment, Epper et al. (2023) find that most altruistic individuals show substantial meritocratic concerns, while selfish individuals tend to exhibit only weak preferences for meritocratic choices. To draw more definitive conclusions on this matter, a study that also elicits meritocratic preferences in the context of performance evaluations appears valuable.

We have investigated the association between social preferences in general and supervisors' rating behavior, and not a ceteris-paribus variation in the extent to which a supervisors cares for the worker's payoff. It will be an interesting topic for future research to exogenously vary social ties between supervisor and evaluated worker. It appears likely that even though – as we have shown – social preferences in general increase the quality of ratings, stronger social ties towards individual workers will hurt the quality of ratings, as they may give rise to favoritism just as claimed by Prendergast and Topel (1996). If this holds, it would suggest that evaluations are optimally conducted by evaluators who are prosocial in general, but have sufficient social distance to the evaluated workers.

Finally, it appears to be important for future work on subjective performance evaluations to endogenize supervisors' information processing costs, both theoretically as well as experimentally. In our experiment, less prosocial supervisors are less diligent in the rating task and, in turn, provide less accurate evaluations despite having easy access to information necessary for belief updating. This suggests that there are costs of processing this information – even in settings where information is directly available – and prosocial agents are more willing to invest them. It seems important to study the robustness of the findings in settings where the costs of information acquisition vary, as this will allow

[42]It appears likely that this effect is even larger in settings where supervisors are aware that their ratings will affect personnel decisions, as prosocials may then additionally take the externalities into account that more accurate information yields (for instance, externalities due to better job assignments).

a deeper understanding of the interplay between preferences, incentives and cognitive costs in the performance evaluation process.

References

Angrist, J.D., Pischke, J.S., 2009. Mostly harmless econometrics: An empiricist's companion. Princeton University Press.

Arechar, A.A., Gächter, S., Molleman, L., 2018. Conducting interactive experiments online. Experimental Economics 21, 99–131.

Arvey, R., Murphy, K., 1998. Performance evaluation in work settings. Annual Review of Psychology 49, 141–168.

Ball, I., 2022. Scoring strategic agents. Mimeo.

Bellemare, C., Sebald, A., 2019. Self-confidence and reactions to subjective performance evaluations. Mimeo.

Berger, J., Harbring, C., Sliwka, D., 2013. Performance appraisals and the impact of forced distribution—an experimental investigation. Management Science 59, 54–68.

Bieleke, M., Dohmen, D., Gollwitzer, P.M., 2020. Effects of social value orientation (SVO) and decision mode on controlled information acquisition—a mouselab perspective. Journal of Experimental Social Psychology 86, 103896.

Bol, J.C., Kramer, S., Maas, V.S., 2016. How control system design affects performance evaluation compression: The role of information accuracy and outcome transparency. Accounting, Organizations and Society 51, 64–73.

Bol, J.C., Smith, S.D., 2011. Spillover effects in subjective performance evaluation: Bias and the asymmetric influence of controllability. Accounting Review 86, 1213–1230.

Bolton, G.E., Kusterer, D.J., Mans, J., 2019. Inflated reputations: Uncertainty, leniency, and moral wiggle room in trader feedback systems. Management Science 65, 4951–5448.

Breuer, K., Nieken, P., Sliwka, D., 2013. Social ties and subjective performance evaluations: an empirical investigation. Review of Managerial Science 7, 141–157.

Bruhin, A., Fehr, E., Schunk, D., 2019. The many faces of human sociality: Uncovering the distribution and stability of social preferences. Journal of the European Economic Association 17, 1025–1069.

Cappelen, A.W., Hole, A.D., Sørensen, E.Ø., Tungodden, B., 2007. The pluralism of fairness ideals: An experimental approach. American Economic Review 97, 818–827.

Cappelen, A.W., Mollerstrom, J., Reme, B.A., Tungodden, B., 2022. A meritocratic origin of egalitarian behavior. Economic Journal 132, 2101–2117.

Casey, L.S., Chandler, J., Levine, A.S., Proctor, A., Strolovitch, D.Z., 2017. Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection. SAGE Open 7, 1–15.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. The quarterly journal of economics 117, 817–869.

Chen, D.L., Schonger, M., Wickens, C., 2016. otree—an open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance 9, 88–97.

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

Dickson, E.S., Gordon, S.C., Huber, G.A., 2009. Enforcement and compliance in an uncertain world: An experimental investigation. Journal of Politics 71, 1357–1378.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., 2009. Homo reciprocans: Survey evidence on behavioural outcomes. Economic Journal 119, 592–612.

Engelmann, D., Strobel, M., 2004. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. American economic review 94, 857–869.

Epper, T., Senn, J., Fehr, E., 2023. Who are the meritocrats? Mimeo, University of Zurich 435.

Fehr, E., Charness, G., 2023. Social preferences: Fundamental characteristics and economic consequences. IZA Discussion Paper No. 16200 .

Frankel, A., Kartik, N., 2019. Muddled information. Journal of Political Economy 127, 1739–1776.

Frankel, A., Kartik, N., 2021. Improving information from manipulable data. Journal of the European Economic Association 20, 79–115.

Gibbons, R., Waldman, M., 1999. A theory of wage and promotion dynamics inside firms. Quarterly Journal of Economics 114, 1321–1358.

Golman, R., Bhatia, S., 2012. Performance evaluation inflation and compression. Accounting, Organizations and Society 37, 534–543.

Gourieroux, C., Monfort, A., 1995. Statistics and econometric models. volume 1. Cambridge University Press.

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

Grosch, K., Rau, H.A., 2017. Gender differences in honesty: The role of social value orientation. Journal of Economic Psychology 62, 258–267.

Horton, J., Rand, D.G., Zeckhauser, R., 2011. The online laboratory: conducting experiments in a real labor market. Experimental Economics 14, 399–425.

Jawahar, I.M., Williams, C.R., 1997. Where all the children are above average: the performance appraisal purpose effect. Personnel Psychology 50, 905–925.

Kampkötter, P., Sliwka, D., 2018. More dispersion, higher bonuses? On differentiation in subjective performance evaluations. Journal of Labor Economics 36, 511–549.

Kane, J.S., Bernardin, H.J., Villanova, P., Peyrefitte, J., 1995. Stability of rater leniency: Three studies. Academy of Management Journal 38, 1036–1051.

Landy, F.J., Farr, J.L., 1983. The Measurement of Work Performance: Methods, Theory, and Applications. Academic Press, New York.

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

Marchegiani, L., Reggiani, T., Rizzolli, M., 2016. Loss averse agents and lenient supervisors in performance appraisal. Journal of Economic Behavior & Organization 131, 183–197.

Markussen, T., Putterman, L., Tyran, J.R., 2016. Judicial error and cooperation. European Economic Review 89, 372–388.

Murphy, K.R., Cleveland, J.N., 1995. Understanding Performance Appraisal. Sage, Thousand Oaks.

Murphy, R.O., Ackermann, K.A., Handgraaf, M.J.J., 2011. Measuring social value orientation. Judgment and Decision Making 6, 771–781.

Ockenfels, A., Sliwka, D., Werner, P., 2015. Bonus payments and reference point violations. Management Science 61, 1496–1513.

Ockenfels, A., Sliwka, D., Werner, P., 2020. Multirater performance evaluations and incentives. Mimeo, University of Cologne.

Paolacci, G., Chandler, J., Ipeirotis, P.G., 2010. Running experiments on amazon mechanical turk. Judgment and Decision Making 5, 411–419.

Prendergast, C., Topel, R., 1996. Favoritism in organizations. Journal of Political Economy 104, 958–978.

Prendergast, C.J., 1999. The provision of incentives in firms. Journal of Economic Literature 37, 7–63.

Rammstedt, B., John, O.P., 2005. Kurzversion des Big Five Inventory (BFI-K): Entwicklung und Validierung eines ökonomischen Inventars zur Erfassung der fünf Faktoren der Persönlichkeit. Diagnostica 51, 195–206.

Rice, S.C., 2012. Reputation and uncertainty in online markets: An experimental study. Information Systems Research 23, 436–452.

Rynes, S.L., Gerhart, B., Parks, L., 2005. Personnel psychology: Performance evaluation and pay for performance. Annual Review of Psychology 56, 571–600.

Schleicher, D.J., Baumann, H.M., Sullivan, D.W., Yim, J., 2019. Evaluating the effectiveness of performance management: A 30-year integrative conceptual review. Journal of Applied Psychology 104, 851.

Sebald, A., Walzl, M., 2014. Subjective performance evaluations and reciprocity in principal-agent relations. Scandinavian Journal of Economics 116, 570–590.

Villeval, M.C., 2020. Feedback policies and peer effects at work, in: Zimmermann, K.F. (Ed.), Handbook of Labor, Human Resources and Population Economics. Springer.

A Appendix

A.1 Proofs

Proof of Proposition 2:
(i) Substituting the optimal rating (2) into the expression for the squared error (1) we obtain

$$\frac{\sigma_a^2\sigma_\varepsilon^2}{n\sigma_a^2+\sigma_\varepsilon^2}+\left(\frac{\eta(\beta+b)}{\gamma+\lambda}\right)^2$$

from which (3) follows as $E\left[\eta^2\right]=V[\eta]+E[\eta]^2$.
(ii) First note that the population $R^2_{a|r}$ is equal to

$$R^2_{a|r}=1-\frac{V[a|r]}{V[a]}.$$

As a and r are jointly normal, we have that the conditional variance of a given r is

$$V[a|r]=V[a]-\frac{(Cov[a,r])^2}{V[r]}$$

such that

$$R^2_{a|r}=1-\frac{V[a]-\frac{(Cov[a,r])^2}{V[r]}}{V[a]}=\frac{(Cov[a,r])^2}{V[a]V[r]} \tag{6}$$

Using that

$$\begin{aligned}
Cov[a,r] &= Cov\left[a,\frac{\eta(\beta+b)}{\gamma+\lambda}+\frac{\sigma_\varepsilon^2 m}{n\sigma_a^2+\sigma_\varepsilon^2}+\frac{n\sigma_a^2}{n\sigma_a^2+\sigma_\varepsilon^2}\left(a+\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right)\right]\\
&= \frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2} \text{ and}\\
V[r] &= V\left[\frac{\eta(\beta+b)}{\gamma+\lambda}+\frac{\sigma_\varepsilon^2 m}{n\sigma_a^2+\sigma_\varepsilon^2}+\frac{n\sigma_a^2}{n\sigma_a^2+\sigma_\varepsilon^2}\left(a+\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right)\right]\\
&= \left(\frac{\beta+b}{\gamma+\lambda}\right)^2\sigma_\eta^2+\frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}
\end{aligned}$$

we obtain

$$R^2_{a|r}=\frac{\left(\frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}\right)^2}{\sigma_a^2\left(\left(\frac{\beta+b}{\gamma+\lambda}\right)^2\sigma_\eta^2+\frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}\right)}$$

which can be rearranged to obtain (4).
(iii) First, note that a Bayesian decision-maker's expectation on a given r is

$$\begin{aligned}\hat{a} &= E[a|r] = m + \frac{Cov[a,r]}{V[r]}(r - E[r]) \\ &= m + \frac{\frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}}{\left(\frac{\beta+b}{\gamma+\lambda}\right)^2\sigma_\eta^2 + \frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}}\left(r - \frac{m_\eta(\beta+b)}{\gamma+\lambda} - m\right)\end{aligned}$$

such that

$$\begin{aligned}V[\hat{a}] &= \left(\frac{\frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}}{\left(\frac{\beta+b}{\gamma+\lambda}\right)^2\sigma_\eta^2 + \frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}}\right)^2 V[r] \text{ and} \\ Cov[a,\hat{a}] &= Cov\left[a, \frac{\frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}}{\left(\frac{\beta+b}{\gamma+\lambda}\right)^2\sigma_\eta^2 + \frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}} r\right] = \frac{\frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}}{\left(\frac{\beta+b}{\gamma+\lambda}\right)^2\sigma_\eta^2 + \frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}} Cov[a,r]\end{aligned}$$

Expected profits from job assignment are

$$m + \Pr(\hat{a} > m)\, E[a - m | \hat{a} - m > 0]$$

which by property B.46 in Gourieroux and Monfort (1995, p. 486) becomes (with $\phi(x)$ being the pdf of a standard normal distribution)

$$\begin{aligned}m + \rho_{a\hat{a}}\sigma_a\phi(0) &= m + \frac{1}{\sqrt{2\pi}}\frac{Cov[a,\hat{a}]}{\sqrt{V[\hat{a}]}} \\ &= m + \frac{1}{\sqrt{2\pi}}\frac{\frac{\frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}}{\left(\frac{\beta+b}{\gamma+\lambda}\right)^2\sigma_\eta^2 + \frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}} Cov[a,r]}{\sqrt{\left(\frac{\frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}}{\left(\frac{\beta+b}{\gamma+\lambda}\right)^2\sigma_\eta^2 + \frac{n\sigma_a^4}{n\sigma_a^2+\sigma_\varepsilon^2}}\right)^2 V[r]}} \\ &= m + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{(Cov[a,r])^2}{V[r]}}\end{aligned}$$

which by using (6) is equivalent (5). ∎

A.2 Effect of signal precision on average ratings

Table A.1: The effect of signal precision on average ratings (treatments with performance pay)

	(1) No acc. inc.	(2) Acc. inc.	(3) Pooled
Signal average	0.612***	0.548***	0.577***
	(0.0755)	(0.0688)	(0.0510)
Four signals	-1.810	-0.683	-1.214
	(2.696)	(2.433)	(1.814)
Accuracy pay			-5.475***
			(1.814)
Constant	26.46***	23.09***	27.60***
	(4.005)	(3.655)	(2.884)
Observations	260	260	520
R^2	0.207	0.233	0.227

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.3 Workers' reactions to ratings

In Part 3, workers learn their actual performance, the rating they received, and whether their bonus depends on the rating or not. Although our analysis in this paper focuses on the supervisor's evaluations and our model does not speak directly to the question of how workers react to their ratings, possible negative reactions to critical feedback are an important reason for leniency frequently mentioned in the subjective performance evaluation literature (see e.g. Golman and Bhatia, 2012, Sebald and Walzl (2014), Ockenfels et al. (2015)). In this section, we explore the hypothesis that receiving a rating below their actual performance triggers a negative reaction of the worker towards the supervisor. To do this we use the worker's willingness to share money with the respective supervisor as measured by the SVO angle elicited again in a series of dictator game choices with the respective supervisor as recipient. A larger SVO angle implies a larger amount given to the supervisor relative to the amount kept for oneself, which we interpret as a kinder reaction of the worker to the rating.

This analysis is related to Sebald and Walzl (2014) and Bellemare and Sebald (2019), who also experimentally study workers' reactions to ratings in a subjective performance evaluation context. In their experiment, Sebald and Walzl (2014) measure workers' beliefs about their performance and find that workers punish supervisors when they are rated

Table A.2: The effect of ratings on workers' propensity to share with the supervisor (SVO angles)

	(1) Perf. inc.	(2) No perf. inc.	(3) Pooled	(4) Pooled
Rating deviation	0.164***	0.131***	0.125***	0.201***
	(0.0315)	(0.0432)	(0.0423)	(0.0532)
Actual performance	-0.0130	0.0292	0.000823	0.00244
	(0.0464)	(0.0586)	(0.0366)	(0.0366)
Rating dev. × Performance pay			0.0422	
			(0.0513)	
Performance pay			1.128	1.169
			(1.158)	(1.146)
max{Rating dev., 0}				-0.0772
				(0.0757)
Constant	19.49***	16.55***	17.75***	18.34***
	(2.159)	(2.589)	(1.777)	(1.864)
Observations	510	254	764	764
R^2	0.057	0.036	0.054	0.055

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable is the worker's SVO angle as measured by a series of dictator game choices by the respective worker with the evaluating supervisor as recipient. *Rating dev.* is the difference between the rating and the actual performance. The inclusion of $max\{Rating\,dev.,0\}$ allows for a difference in the slope of the rating deviation when the rating exceeds the actual performance. Data in (1) from treatments P-NA-S1, P-A-S1, P-NA-S4, P-A-S4, in (2) from treatments NP-A-S1 and NP-NA-S1, and in (3) from all treatments.

below their perceived performance.[43]

We find that when workers receive a rating below their performance, their average SVO is 15.9, while it is 21.7 when the rating is above their performance ($p \leq 0.001$, two-sided t-test). In terms of the respective monetary impact, "underrated" workers on average give up 23 cents to increase their supervisor's payoff by 42 cents, while "overrated" workers give up 33 cents to increase their supervisor's payoff by 59 cents.

Table A.2 reports regressions of the workers' propensity to reciprocate the rating as measured by the SVO on the deviation between the rating and the workers' actual performance controlling for actual performance. As by design the rating deviation is exogenously assigned conditional on the rating (through random assignment of supervisors and their signals to workers), it estimates the causal effect of the rating on the workers' propensity to reciprocate. We run separate regressions for treatments with and without performance pay (given that this is information that workers receive) and for the pooled data. As the regression results show, workers indeed reciprocate higher ratings. Importantly, they do so not only when they materially benefit from the rating (column (1)) but also when the rating has no material consequences for the workers (column (2)). We find no significant difference in the extent to which workers reciprocate ratings with and without performance pay (column (3)).

In column (4) we explore whether the reciprocal reaction is driven rather by punishing those that "underrate" or rewarding those that "overrate" actual performance. We do so by additionally including a variable $max\{Rating\ deviation, 0\}$ which allows for a difference in the slope of the reaction function between the ratings above as compared to the ratings below the actual performance. The slope of 0.201 is larger for ratings below the actual performance than the slope of $0.201 - 0.077 = 0.124$ for ratings above. But even the slope above the actual performance is significantly different from zero ($p = 0.001$). Hence, workers not only punish evaluations below their actual performance but they also reward evaluations exceeding it.

Taken together, we find evidence that workers react to the difference between their rating and their *actual* performance, and they do so even when no monetary payments are tied to the rating. This complements results from earlier work (e.g. Sebald and Walzl, 2014 and Bellemare and Sebald, 2019) who find that workers react negatively when they are rated below their *perceived* performance and strengthens the argument made in the literature that anticipated reciprocal reactions from agents can be a source for supervisor rating leniency.

[43]Bellemare and Sebald (2019) extend the experimental setup of Sebald and Walzl (2014) by allowing workers to reward supervisors in addition to punish them. They find that over- and underconfident workers react differently to being rated above and below their belief: Underconfident workers reward being overrated but do not punish supervisors who underrate them, while overconfident supervisors do not react to being overrated but punish supervisors who rate them below their belief about performance.