I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15496

Social Preferences and Rating Biases in Subjective Performance Evaluations

David Kusterer
Dirk Sliwka

AUGUST 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15496

Social Preferences and Rating Biases in Subjective Performance Evaluations

David Kusterer
University of Cologne

Dirk Sliwka
University of Cologne, CEPR, CESifo and IZA

AUGUST 2022

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Social Preferences and Rating Biases in Subjective Performance Evaluations*

We study the determinants of biases in subjective performance evaluations in an MTurk experiment to test the implications of a standard formal framework of rational subjective evaluations. In the experiment, subjects in the role of workers work on a real effort task. Subjects in the role of supervisors observe subsamples 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. In line with the predictions of the model of optimal evaluations we find that ratings are more lenient and less accurate when they determine bonus payments and that rewards for accuracy reduce leniency. When supervisors have access to more detailed performance information their ratings vary to a stronger extent with observed performance. In contrast to the model's prediction we do not find that more prosocial supervisors always provide more lenient ratings, but that they invest more time in the rating task and achieve a higher rating accuracy.

JEL Classification: J33, C91, M52

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

Corresponding author:
Dirk Sliwka
University of Cologne
Faculty of Management, Economics and Social Sciences
Albertus-Magnus-Platz
50923 Köln
Germany
E-mail: sliwka@wiso.uni-koeln.de

* 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, Felix Kölle, Johannes Mans, Frank Moers, Patrick Schmitz, participants of the 2021 annual meeting of the German Association for Experimental Economics in Magdeburg, of the Colloquium on Personnel Economics 2022 in Aarhus and of the Maastricht Behavioral Economic Policy Symposium 2022 for helpful comments and discussions. Moreover, we would like to thank Mary Wack and Niklas Wagner for excellent research assistance.

1 Introduction

When assessing the performance of employees, firms often have to rely on subjective performance evaluations. As in many jobs employees' performance cannot easily be assessed by objective key figures it is common that supervisors are asked to rate the performance of their subordinate employees. It is well known that such evaluations tend to be "biased", that is the distribution of observed subjective assessments deviates systematically from the expected underlying distribution of performance.

It has, for instance, often been observed that ratings tend to be too lenient and too compressed (see e.g. Murphy and Cleveland (1995), Prendergast and Topel (1996), Prendergast (1999)). That is, 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 different mechanisms. For one, as performance evaluations are typically carried out by supervisors that are not owners of the organization, these supervisors' *preferences* may not be aligned with the preferences of the employer. Supervisors thus may willingly deviate from reporting accurate assessments, for instance when they exhibit social preferences towards the employee and have to trade off the employer's against the assessed employee's interests. But inaccurate ratings may also be due to a lack of *information.* That is, even when supervisors care about the accuracy of their ratings, they have incomplete information about the employee's true performance. Finally, observed deviations between true performance and performance assessments may also be due to *cognitive limitations*. That is, supervisors may not be able to process the available information in a fully rational manner which may cause inaccurate evaluations.

Firms use subjective performance evaluations for multiple reasons (see e.g. Landy and Farr (1983); Arvey and Murphy (1998); Prendergast (1999)). For instance, firms use subjective assessments for the allocation of individual bonuses in incentive schemes when objective performance measures are unavailable. Moreover, evaluations are used for personnel decisions such as employee promotions or terminations and to provide direct feedback to employees about their performance. The aforementioned biases affect the usefulness of ratings for these purposes. But supervisors' preferences and, in turn, resulting biases can also be affected by the purpose of the evaluation.

The aim of this paper is to use a standard framework to formally model subjective evaluations by a rational decision maker based on Prendergast and Topel (1996) and test its implications in an experiment.[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 to provide an optimal rating given the noisy information – against potential social preferences towards the

[1]The framework is also applied and extended for instance in Golman and Bhatia (2012), Kampkötter and Sliwka (2018), Manthei and Sliwka (2019).

assessed worker – which implies a tendency for more lenient and, in turn, less accurate ratings.

To test the predictions of the model, we conducted an online experiment on Amazon MTurk, a website for crowdsourced labor.[2] In the experiment, 780 subjects worked on a real-effort task where they had to enter text contained in hard-to-read images (similar to so-called "captchas"). In a next stage, the performance of each of these subjects was evaluated by another set of subjects in the role of supervisors. Supervisors had noisy information on workers' performance as they could gather information only on a randomly drawn subset of the workers' performance outcomes. We implemented six different treatments varying (i) whether workers are paid according to the rating or not, (ii) whether supervisors are paid according to accuracy or not, and (iii) whether supervisors observe a larger or smaller subset of the workers' performance outcomes.

We derive hypotheses for each of these treatment variations based on the framework. We find that the observed patterns are mostly well-organized by the formal model. First of all, ratings are significantly higher for the same performance when workers are paid according to the ratings.[3] That is, supervisors indeed internalize the effect of their ratings on a worker's well-being even in an anonymous experiment where there is no future interaction. Second, paying supervisors for the accuracy of their rating reduces the rating leniency triggered by worker bonuses. Third, when supervisors can access a higher number of available signals their ratings vary more with observed performance.

We also investigate the effects of the treatments on the average rating error, i.e. the squared deviation between true and reported performance. In line with the predictions of the model, performance pay for workers reduces the accuracy of ratings (i.e. increases rating error), while paying supervisors for accuracy and providing more performance information increases rating accuracy. Hence, the results support the notion that there is indeed a tension between different uses of performance ratings: when ratings are used to reward employees they are less valuable to assess their performance accurately.

In two domains, however, the empirical results deviate from the model's predictions. The model predicts that supervisors with stronger social preferences provide more lenient evaluations, in particular when the rating determines a bonus for the assessed worker. In the experiment we assess each supervisor's social preferences through the incentivized Social Value Orientation (SVO) measure (Murphy et al. (2011)). The SVO measure consists of several dictator games where subjects choose an allocation of money between

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

[3]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.

themselves and a randomly matched worker. In contrast to the hypothesis, we do not find systematic evidence that supervisors who exhibit stronger social preferences provide more lenient ratings. Exploring reasons for this observation, we find that stronger social preferences are also associated with a *tendency to take the rating task more seriously* and rate more accurately. For instance, we find that supervisors classified as prosocial rather than individualistic by the SVO take significantly more time for the rating (+28%), their ratings follow observed signals to a stronger extent, and they tend to exhibit lower rating errors. Hence, our results show that supervisors' social preferences do not necessarily undermine the precision of ratings due to rating leniency as the model predicted but may rather lead to more accurate rating behavior.

Moreover, in contrast to the model's predictions we do not find that the higher accuracy induced by access to more performance signals leads to a higher rating variance. While with more precise information ratings vary to a stronger extent with observed signals (which increases rating variance and is in line with the model), rating errors due to suboptimal information processing are reduced (which decreases rating variance and is not in line with the model as supervisors in the model already correctly update beliefs according to Bayesian rule). That is, behavior is closer to the prediction implied by rational Bayesian updating when there are more available signals. A potential explanation is that due to economizing cognitive costs of information processing supervisors invest more in belief updating when it is more worthwhile to do so (compare for instance Kominers et al. (2018) for a formal analysis of costly Bayesian inference).

We contribute to the literature on subjective performance evaluations in several respects. First, we show that a formal framework based on Prendergast and Topel (1996) organizes the experimental data in many dimensions quite well. The previous 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. 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 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. We also find that workers punish low ratings and reward high ratings (compared to 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 adverse reaction by the worker (see Golman and Bhatia, 2012 or Ockenfels et al., 2015). However, we show that rating leniency also occurs when supervisors do not have to anticipate these reactions as long as they know that their ratings affect workers' pay.

Our results showing that a stronger rating differentiation with respect to performance can go along with a lower rating variance due to a reduction in rating mistakes also calls for some caution when using rating variance as a measure of rating differentiation with respect to performance, which sometimes has been done in field studies on performance ratings (see e.g. Bol (2011), Engellandt and Riphahn (2011), or Kampkötter and Sliwka (2018)). However, as those papers tend to find a positive association between rating variance and subsequent performance, our results may indicate that they may even have underestimated the effects of performance-based rating differentiation.

A recent tool used by organizationsto counteract biased ratings are calibration committees, where groups of supervisors assign ratings 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 such that e.g. calibration committees may be less beneficial under such circumstances.

With respect to uncertainty in 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). 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. 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 behavior when rating performance.

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

[4]In their experiment 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 punished by receiving lower performance evaluations themselves, while supervisors who give less compressed ratings are more likely to receive a promotion.

mental results and Section 5 concludes.

2 A Simple Model

2.1 The Setup

Consider the following simple framework which builds on Prendergast and Topel (1996) to model subjective performance evaluations.[6] A supervisor evaluates the performance 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.[7] 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 higher rating (i.e. $b > 0$) such that her utility is

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

The supervisor may have social preferences and therefore her utility may be affected by the well-being of the agent. As in Prendergast and Topel (1996), 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 accuracy and λ determines her material incentives to provide accurate ratings.[8]

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 more performance signals.

[6] This framework is, for instance, also 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.

[7] Note that we focus on the evaluation task and thus do not model an endogenous effort choice.

[8] In Prendergast and Topel (1996), 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.

2.2 Optimal Evaluations and Hypotheses

First, note that the signal average $\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i$ is a sufficient statistic for estimating a, i.e. $E\left[(r-a)^2\middle|s\right] = E\left[(r-a)^2\middle|\bar{s}\right]$. The supervisors' decision problem when choosing the 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[9]

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

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*

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

Supervisors in our experiment may of course differ in their social preferences η. To derive predictions on the population of supervisors, assume that their social preference parameters follow a normal distribution $\eta \sim N\left(m_\eta, \sigma_\eta^2\right)$. Expected ratings are then simply obtained by replacing η with m_η in equation (2).

Proposition 1 has several implications that we test in the experiment. 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 the agents receive a bonus which depends on the performance rating, supervisors with social preferences will internalize the effect on the agents' well-being to some extent – thus shifting ratings upwards. However, this will not affect the marginal effect of the observed signal on reported performance (the slope of the rating function). Thus the bonus will not affect rating compression.

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 ($\beta > 0$).*

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

The introduction of a reward for accuracy increases the costs of rating leniency. As leniency thus becomes more costly for the supervisor, this will lead to less lenient ratings. Moreover, as ratings are more lenient when there is performance pay, the reduction in leniency is also larger in this case. To see that, note the following: When the intrinsic private benefit b from higher ratings is small and there is no bonus, there will be hardly any rating leniency in the first place such that supervisors would prefer accurate ratings even without being rewarded for accuracy. If, however, there is performance pay, the urge to deviate from accurate ratings is larger and thus material incentives for accuracy become more important.

We also vary the number of signals observed by the supervisor in the experiment (see Ockenfels et al. (2020) for a similar analysis):

Hypothesis 3: *If the supervisor observes more signals, performance 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 achieves a more precise estimate of the agent's true performance and therefore is willing to deviate more from her prior expectation m about performance. That is, low signals will lead to lower assessments and high signals to higher assessments and overall rating compression decreases.

The model also predicts that – while rating compression decreases – 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), 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 Golman and Bhatia 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.[10] With less precise signals, errors are more likely and hence the supervisor shifts ratings upwards . When signals become more precise this upwards 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 framework presented here.

The model also makes a prediction on the role of social preferences η as the optimal rating r is increasing in η:

[10] To be precise, the key difference between their model and framework analyzed here is that (besides using a linear instead of a quadratic functional form) 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 $\lambda > 1$.

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 others, she will prefer to assign higher ratings. This will in particular be the case when bonus payments are in place as here higher ratings create a material benefit for the assessed agent. We caution, however, that this hypothesis cannot be tested in a similarly clean causal manner as the previous hypotheses: while the assignment to the social preference type of the supervisor is exogenous in our experiment, any measure of social preferences maybe correlated with other unobserved traits that also may affect rating behavior.

In a next step, we consider how the different parameters we vary affect the precision of the ratings. To do so, we consider the expected (squared) rating error, i.e. the expected deviation between rating and true performance.

Proposition 2. *The expected squared rating error is*

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

Proof. 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 $E\left[\eta^2\right] = V[\eta] + E[\eta]^2$. □

This result implies the following hypothesis:

Hypothesis 5: *Rating errors are larger when agents receive bonus payments ($\beta > 0$) and smaller when supervisors are rewarded for accuracy ($\lambda > 0$) and when they observe more performance signals. A reward for accuracy reduces rating errors to a larger extent when ratings determine agents' bonus payments.*

As already shown by Prendergast and Topel (1996), rating errors are larger when the agents receive bonus payments for two reasons: As we have seen before, bonus payments for the agents induce rating leniency and this leads to an upwards bias. But moreover, as the supervisors' social preferences are imperfectly known and supervisors with higher prosocial motives are more prone to rating leniently (as $\partial r/\partial\eta > 0$ in our model), additional noise is added to the rating as a signal of performance. Rating errors are lower when supervisors are rewarded for accuracy as such a reward reduces the leniency bias. As

bonus payments aggravate the leniency effect, rewards for accuracy are more valuable to reduce rating errors when bonuses are used. Finally, rating errors decrease in the number of signals available as a higher number of signals facilitates the accurate rating of performance.

3 Experimental Design

Our experiment consists of three parts. Each part is completed before the next part begins. In Part 1, subjects (called workers) work on a real effort task. After Part 1 is completed, another set of subjects (called supervisors) receives noisy information about the performance of one of the workers from Part 1 and submits a rating about their work.[11] After Part 2 is completed, subjects from Part 1 are invited again to learn their rating. We first describe the tasks in the three parts and go into more detail about the payoffs for workers and supervisors as well the determination of the signal for the supervisor in the description of the treatments below.[12]

3.1 Tasks

Part 1

Workers perform a real-effort task which we call the Entry Task. The Entry Task consists of entering text contained in hard-to-read images (similar to so-called "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. 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. 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. After the Entry Task is finished, we elicit the workers' beliefs about their performance on all 10 pages. 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 they will receive a payment which may depend on the rating they receive. Workers finish this part by filling in a demographics questionnaire.

[11] In the experimental instructions and interface, we refer to all subjects as *workers*, the term also used by MTurk.

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

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 the work of one worker from the first part. Subjects who participated in Part 1 are excluded from participating in Part 2. One worker from Part 1 is randomly matched to each supervisor in Part 2 (described in more detail below). Matching was anonymous and participants never received information on the identity of other subjects.

At the beginning of Part 2, supervisors see the practice page and – in order to familiarize them with the task – work on two pages of the Entry Task. 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 when they are asked to submit a rating of the worker's performance. We also elicit supervisors' beliefs about their own performance on the two example pages.

In the Rating Task (see Figure 1 for a screenshot), supervisors receive a signal about the number of correctly entered images by the matched worker and are asked to rate the worker using an integer rating $r \in [0\%, 100\%]$. Supervisors are shown the number of correctly entered images from a randomly drawn subset of the 10 pages the worker filled in (the subsets are described in greater detail in the description of the treatments below). Supervisors also see a histogram of all workers' average performances (over all 10 screens) as well as the mean and the standard deviation. For their matched worker, they also see a table with rows for each page of the Entry Task. For pages in the signal subset, the table contains the number of correctly entered images, while for pages not in the subset no information is given. They are told that the rating should reflect the worker's performance on all 10 pages of the Entry Task.

After the Rating Task, supervisors complete the incentivized Social Value Orientation (SVO) slider measure with a randomly selected worker (but not the one they rated) as the recipient in order to measure their social preferences towards the worker population. The SVO slider measure consists of several dictator games where a subject can choose an allocation of money between themselves and a matched partner. We use the six primary items from Murphy et al. (2011) where each point corresponds to \$0.01 (see Appendix A.1 for more details). Supervisors also fill in a 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. At the end of Part 2, supervisors learn their total payment consisting of the payments from their matched worker's performance, from their rating accuracy (depending on the treatment) and from their choices in the SVO task.

Table 1: Treatments

Name	γ	β	n
NP-NA-S1	0	\$0	1
NP-A-S1	0.004	\$0	1
P-NA-S1	0	\$2	1
P-A-S1	0.004	\$2	1
P-NA-S4	0	\$2	4
P-A-S4	0.004	\$2	4

Part 3

Workers who completed Part 1 and received a rating in Part 2 are invited per email to participate in Part 3. First, they learn their rating and their actual average performance and are asked to state both their satisfaction with their performance and their satisfaction with the rating on Likert scales from 0-10. Second, they learn whether their payment depends on the rating, and learn their payment. They complete the incentivized SVO slider measure with their supervisor as the recipient to measure their social preferences towards their supervisor as a reaction to their rating and the associated payment (see Murphy et al., 2011). After this, they learn their total payment consisting of payments from the Entry Task, from the SVO task they completed, and from the SVO task a different supervisor than the one who rated their performance completed with them as the recipient in Part 2. Additionally, supervisors receive the payment from their workers' choices in the SVO task. This concludes the experiment.

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 P and A with low signal precision, S1. For high signal precision, S4, workers are always paid according to the rating and we only vary A. Altogether, we conduct 6 treatments (see Table 1). We randomly assign workers (and hence matched supervisors) to the six treatments, stratifying the assignment to obtain similar performance distributions across treatments.

The supervisors receive a monetary payoff of $\$4 - \lambda(r-a)^2 + 0.01a$. The constant $\lambda \in \{0, 0.004\}$ determines whether supervisors are paid according to the accuracy of their rating. The positive value of λ was chosen such that the supervisor's payoff is nonnegative even for the lowest rating accuracy. Supervisors also receive \$0.01 for each image their

Rating Task

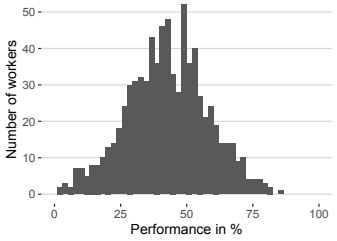
Show Instructions

Distribution of performance

The **average performance** of 780 workers who completed the Entry Task is **42.5%**. This means that on average, they entered the text correctly on 42.5 of the 100 images. The **standard deviation** (a measure of how far the performance is spread out) of these workers is **15.5**.

The graph below shows the distribution of performance of 780 workers who completed the Entry Task.

You can read it in the following way: For each level of performance (on the axis at the bottom), it shows the number of workers with that performance (on the axis at the left).



Your worker's performance

The worker matched to you had the following performance on 1 out of 10 pages that was randomly selected. Note that their performance on the other 9 pages will not be revealed to you.

Remember that the 10 pages had different time limits such that the revealed performance can be from a page with any of the time limits mentioned in the Instructions (17, 19, 21, 23, or 25 seconds).

Page	Number of correct entries
1	
2	
3	**5**
4	
5	
6	
7	
8	
9	
10	

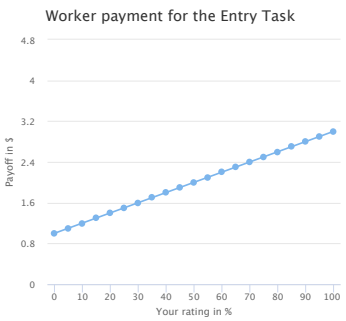
The payoffs

Your payment:

- For the Rating Task you receive $4.00 − 0.9 × (true performance - rating)^2/2250, but not less than $0.00.
  The payment will be the higher, the closer your rating is to the true performance (see figure below).
- You will also receive $0.01 for every image the worker matched to you entered correctly over all 10 pages.
  For example, if they entered 0 images correctly, you receive $0.00, if they entered 50 images correctly, you receive $0.50, and if they entered 100 images correctly, you receive $1.00. (This payment does not depend on your rating, only on the worker's actual performance.)

The worker receives a payment of $1.00 + $2.00 × (**your rating**)/100.
The worker's payment increases in your rating (see figure below). The higher the rating you give, the higher the worker's payment will be. (The worker's payment is paid by us and not deducted from your earnings.)

These graphs illustrate your payment for the Rating Task and the payment of the worker matched to you based on the rating you give and their true performance:



Your rating

How would you rate the worker?
As a guidance, ratings should reflect the percentage of correctly entered images by the worker.

%

Given your currently entered rating, the worker would receive a bonus of $--.

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

matched worker entered correctly. The monetary payment to the worker is $\$1 + \beta \frac{r}{100}$. The constant $\beta \in \{\$0, \$2\}$ determines whether the workers is paid according to the supervisor's rating. Supervisors are informed about both their own and the worker's payoff function. In case the supervisor's payoff depends on accuracy, supervisors are shown a plot of their payoff as a function of their accuracy. 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 (see 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 signal subsets the supervisor saw or whether the supervisor's payment depends on accuracy. 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.

3.3 Procedures

We conducted the experiment online on Amazon MTurk, a website for crowdsourced labor.[13] 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.[14] 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.[15] 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,[16] 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[17] 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

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

[14] 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.

[15] The threshold of 1000 HITs does not imply that a worker has participated in 1000 experiments. Most HITs on MTurk contain repetitive small tasks like classifying an image (each one a single HIT) which can be worked on in batches such that new workers can reach the number of 1000 HITs in a matter of weeks.

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

[17] 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.

of the instructions, and that if they do not answer a question correctly after the third attempt, they will be excluded from further participation. Subjects could review the instructions during answering the comprehension questions. Subjects in Part 1 are 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. Subjects received a reward of \$0.50 for completing a part of the experiment, i.e., even if a worker did not show up again for Part 3, they received the reward for completing Part 1.

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. This was ensured by using MTurk qualifications, labels that can be given to a worker and checked against in order to accept or reject workers from participating in a study. 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. Using a time limit on MTurk is necessary as it is common for workers to accept tasks without starting to work on them, blocking slots for workers who are directly available and thus delaying the experiment. 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.[18] After we gathered a rating for each worker in Part 2, we emailed reminders to the workers from Part 1 to participate in the final part to learn their rating. In Part 3, 764 of 780 subjects from Part 1 returned. The attrition rate of 2.1% did not differ significantly between the treatments as workers only learned about the treatments in Part 3.

As to the demographics of our sample, 50.5% of our subjects are female. On average, they are 37.9 years old, 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.

[18]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

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 the six treatments we conducted. 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 treatment both in terms of means and variances. We explore these treatment differences and their drivers in detail in the following sections where we test the four key hypotheses that are implied by Propositions 1 on the rating behavior and Proposition 2 on the resulting accuracy of the ratings.

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 relative to this benchmark setting testing Hypothesis 1, which states that performance pay should lead to stronger rating leniency. To do this we compare treatment NP-NA-S1 with the treatment P-NA-S1 in which agents receive performance pay based

on the supervisors' evaluation. By equation (2) the model predicts that

$$\frac{\partial r}{\partial \beta} > 0 \text{ and } \frac{\partial^2 r}{\partial \bar{s} \partial \beta} = 0. \tag{4}$$

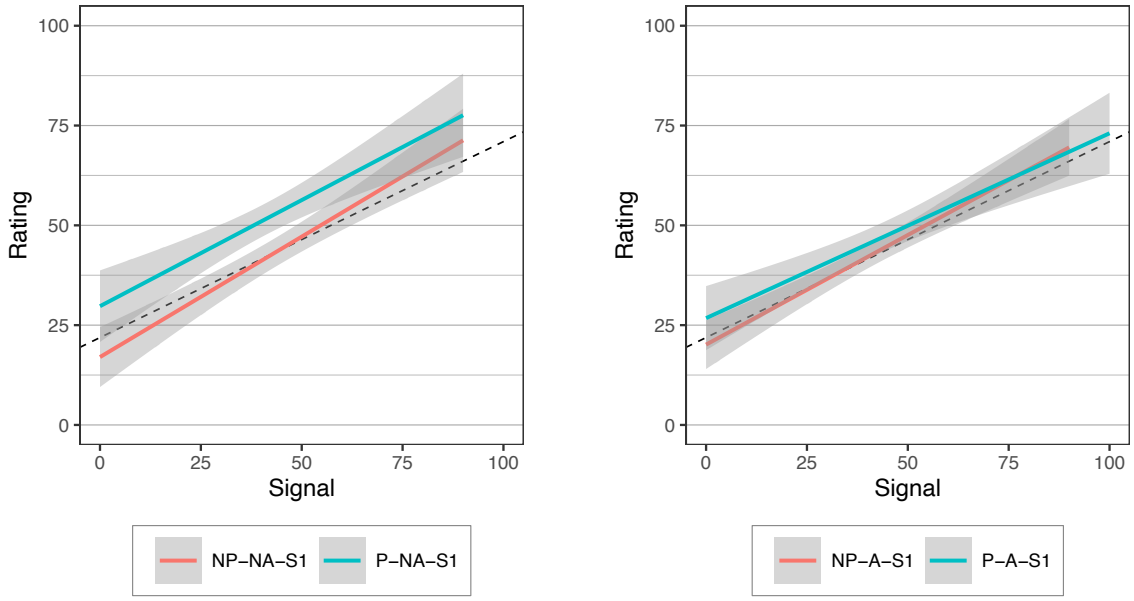
That is, the rating should be higher when performance pay is in place, but the slope of the rating function with respect to the signal should be unaffected. Table 3 reports regressions of the rating on a dummy for the use of performance pay for the agent (column (1)), as well as the signal observed by the supervisor (column (2)), and an interaction term between both (column (3)). The experimental results are well in line with Hypothesis 1. Indeed supervisors become substantially more lenient in their ratings when performance pay is in place. 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 compression: The interaction term of signal and performance pay in column (3) is small and not significantly different from zero.

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

	(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.

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 but does not affect the slope. Taken together, we find evidence



(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 ($\eta = 0$).

for Hypothesis 1 as the introduction to performance pay leads to more lenient ratings.

4.3 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. Here the model implies that providing incentives for accuracy should reduce ratings and that this effect should be stronger when performance pay is in place:

$$\frac{\partial r}{\partial \lambda} < 0 \text{ and } \frac{\partial^2 r}{\partial \lambda \partial \beta} < 0.$$

It is, however, 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

(columns (1) and (2) of Table 4). Indeed, 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 substantially when supervisors' pay depends on the accuracy of their rating. About 2/3 of the leniency effect disappears in this case. As in column (1), we find no evidence that accuracy pay affects rating leniency when there is no performance pay, which again indicates that supervisors may have little concern for potential non-monetary intrinsic private benefits b of ratings. This 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 reveals the downward shift caused by the introduction of accuracy pay when performance pay is in place. Taken together, we do not find evidence for the strong formulation of Hypothesis 2 that accuracy pay always reduces ratings. But, we find support for the second part of the hypothesis that accuracy pay reduces ratings when agents' payments are tied to the ratings.

Note that also in line with the model, as column (4) shows there is no evidence that accuracy pay or performance pay affect the slope (i.e. $\frac{\partial r}{\partial s}$ or the extent to which the rating depends on the observed signal).

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.

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 here implies a shift in both, the intercept of ratings as well as the slope of the rating function as

$$\frac{\partial^2 r}{\partial \bar{s} \partial n} > 0 \text{ and } \left.\frac{\partial r}{\partial n}\right|_{\bar{s}=0} < 0$$

such that performance differentiation increases. That is, the rating should vary with the observed signal (i.e. the average performance across the observed screens) to a stronger extent. As the intercept of the optimal rating function is decreasing in n, low signals should lead to lower evaluations in the S4 treatments than in the S1 treatments and higher signals should lead to higher evaluations.

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 indeed decreases 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, performance differentiation of ratings increases. The effects of the treatment dummy and its interaction with the signal are significant in all models with the exception of the intercept in the model restricted to data without accuracy incentives, where the point estimates are of very similar magnitude.

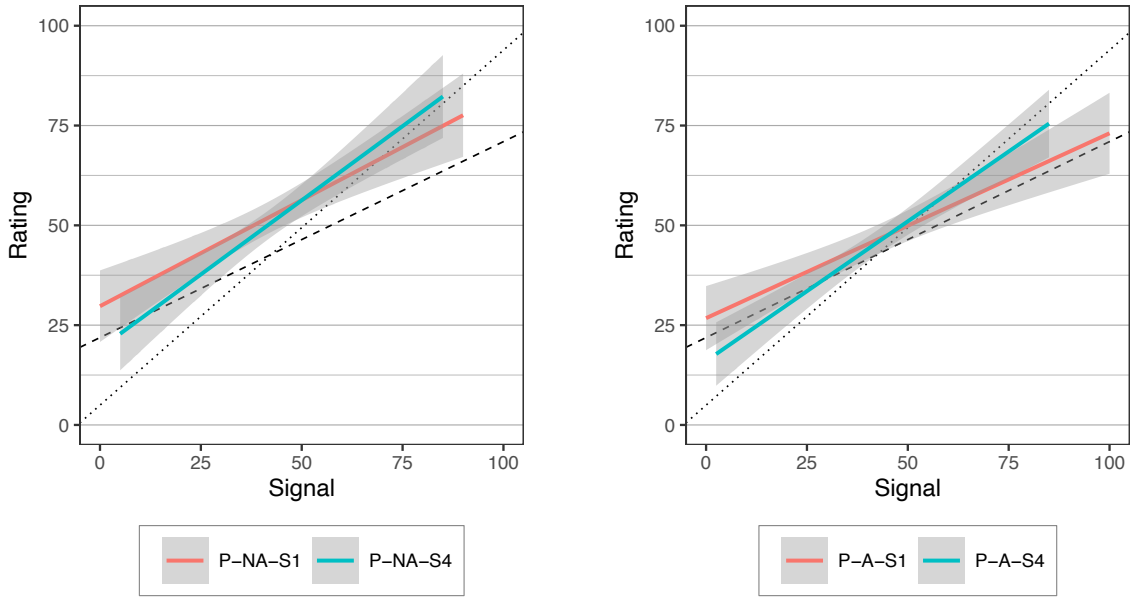
These results thus support Hypothesis 3 and are in line with similar experimental findings in Ockenfels et al. (2020). We also present these results graphically in Figure 3, which mirrors Figure 2 and adds the optimal rating function for supervisors without social preferences (η =0) with four signals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

Table 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.

So far, we have investigated the effect having access to more precise information on the shape of the supervisors' rating function, i.e. conditional on the observed signal. We have seen that, in line with the theoretical model, more information "rotates" the optimal rating function upwards, inducing a lower intercept but a larger slope such that signals have a stronger impact on ratings. But it is also instructive to investigate the effect of the additional information on the average rating (unconditional on the realized



(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)

signal). As noted above, the framework presented in Section 2 predicts that the precision of the available information does not affect overall rating leniency and thus the average rating should not differ between the treatments with one or 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,[19] more precise signals should also reduce rating leniency.[20]

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 the average ratings are larger in the treatments with a less precise signal as predicted by Golman

[19]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.

[20]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).

and Bhatia's (2012) model, the differences are not significant (cf. the regression results in Table A.1 in the Appendix) and relatively small. 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 strong evidence of uncertainty increasing average ratings in our experimental setting. Supervisors with more precise signals tend to rate less leniently, but the effect size is small and not statistically significant.

4.5 How do the Treatments Affect the Accuracy of Ratings?

Moving away from the ratings as the dependent variable, in this section we analyze the effect of our treatments on the accuracy of the ratings measured by the squared deviation from the actual performance. According to Hypothesis 5 both accuracy pay and greater signal precision should decrease rating errors as

$$\frac{\partial E[(r-a)^2]}{\partial \lambda} < 0 \text{ and } \frac{\partial E[(r-a)^2]}{\partial n} < 0,$$

while performance pay should lead to increased rating errors as

$$\frac{\partial E[(r-a)^2]}{\partial \beta} > 0.$$

Figure 4 shows the average squared rating errors for all treatments. We see the largest rating errors when there is only one signal and there is no reward for accuracy but performance pay is used. The treatment variables have the expected directional effects: Introducing accuracy pay, adding more signals and removing performance pay all decrease rating errors.

To test Hypothesis 5, we report two regressions with the squared rating error as the dependent variable in Table 6. The model in column (1) confirms the first part of Hypothesis 5: Ceteris paribus, performance pay increases rating errors while accuracy pay and access to more signals leads to more accurate ratings.

Equation 3 also predicts an interaction between performance and accuracy pay such that accuracy pay decreases rating errors to a stronger extent when performance pay is in place ($\frac{\partial^2 E[(r-a)^2]}{\partial \lambda \partial \beta} < 0$). The interaction term in column (2) is negative as expected but not significant such that we find no strong evidence for the second part of Hypothesis 5.

Equation 3 also illustrates the importance of heterogeneity in supervisor social preferences for the expected squared rating error. We analyze the impact of social preferences measured by supervisor's Social Value Orientation in greater detail in the following section.

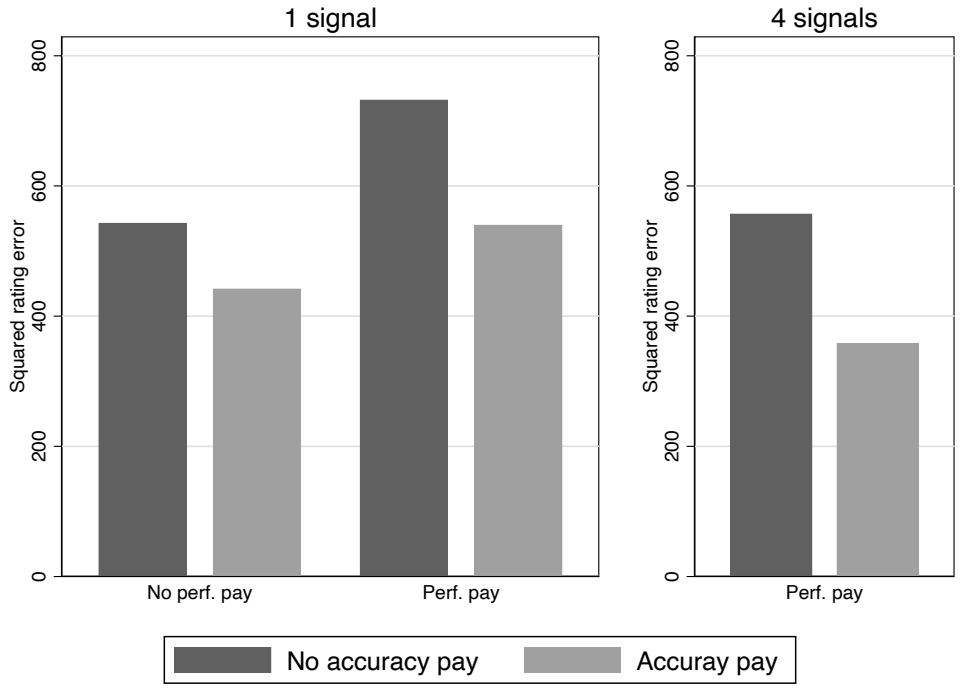


Figure 4: Average squared rating error within treatments.

4.6 Supervisors' Social Preferences

According to the predicted rating function (2), 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, both effects should be less pronounced when the supervisor receives a bonus for accurate ratings as in this case, the supervisor's behavior should vary less with her social preferences.

We can make use of the exogenous variation of the supervisor assignment to test whether there is a correlation between the supervisor's social preference type and the assigned ratings.[21] To measure the supervisors' social preferences towards the worker population (or η in the language of our model) we elicited the their Social Value Orientation (Murphy et al., 2011) and classify supervisors into prosocial and individualistic types.[22]

Table 7 shows regressions of the rating on a dummy indicating whether the supervi-

[21] 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.

[22] The SVO elicits supervisor's choices in a series of dictator games. For more details on the classification see Appendix A.1.

Table 6: The effects of performance pay, accuracy pay, and signal precision on squared rating error

	(1)	(2)
Performance pay	143.6*	190.9*
	(78.43)	(103.7)
Accuracy pay	-164.1**	-101.0
	(64.07)	(92.03)
Four signals	-178.4**	-178.4**
	(84.36)	(84.40)
Performance pay × Accuracy pay		-94.59
		(124.9)
Constant	574.7***	543.2***
	(57.71)	(69.24)
Observations	780	780
R^2	0.016	0.016

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. *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.

sor is prosocial interacted with a dummy for the use of performance pay.[23] Column (1) considers the treatments without incentives for accuray (i.e. treatments NP-NA-S1 and P-NA-S1). We find that prosocial supervisors indeed provide significantly more lenient ratings when there is no performance pay. Interestingly, we do not find that this effect becomes stronger when there is performance pay (and no incentives for accuracy): Contrary to the prediction from our model (i.e. that $\frac{\partial r}{\partial \eta \partial \beta} > 0$), we even observe rather the opposite pattern as the interaction term is negative[24] and the sum of the *Prosocial* dummy and the interaction term is not significantly different from zero. To be clear, the ratings of prosocial supervisors still become more lenient when performance pay is introduced, but the gap between ratings from prosocial and individualistic supervisors tends to vanish when workers are paid according to the rating.

Column (2) shows the results for the treatments where the supervisor is rewarded for accuracy (i.e. NP-A-S1 and P-A-S1). In line with the hypothesis that prosocial preferences should affect ratings to a lesser extent under accuracy incentives we here do not find sizeable differences between prosocial and individualistic supervisors.

The results thus show the notable pattern that performance pay leads to more leniency

[23]Figure A.2 in the Appendix shows average ratings and average squared rating errors for all treatments and conditional on SVO type.

[24]Note that it is unlikely that this reflects ceiling effects as the actual ratings are substantially smaller than 100, the upper bound of the rating scale.

Table 7: The effect of SVO type on ratings

	(1) No acc. inc.	(2) Acc. inc.
Prosocial	8.737**	0.334
	(3.633)	(2.936)
Performance pay	12.71***	5.947
	(4.294)	(3.959)
Prosocial × Performance pay	-5.956	-4.867
	(5.457)	(4.929)
Signal average	0.569***	0.505***
	(0.0636)	(0.0564)
Constant	13.93***	21.73***
	(3.312)	(3.022)
Observations	260	260
R^2	0.290	0.270

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable is the assigned rating. *Prosocial* is a dummy variable indicating whether the supervisor is classified as prosocial based on the SVO measure. *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 in (1) from treatments NP-NA-S1 and P-NA-S1 and in (2) from NP-A-S1 and P-A-S1.

in general but this increased leniency is not driven by prosocial supervisors but to the contrary appears to be rather driven by the the less prosocial ones.

Table 8: The effect of SVO type on rating accuracy and squared rating errors

	(1) Rating	(2) Sq. rating error
Prosocial	-6.627*	-116.6*
	(3.548)	(66.17)
Signal average	0.482***	
	(0.0610)	
Prosocial × Signal average	0.175**	
	(0.0755)	
NP-A-S1	0.808	-92.93
	(2.319)	(92.51)
P-A-S1	3.720	5.971
	(2.574)	(113.2)
P-A-S4	3.150	-189.3**
	(2.373)	(95.24)
P-NA-S1	9.563***	194.6*
	(2.668)	(114.9)
P-NA-S4	8.067***	15.07
	(2.563)	(113.4)
Constant	21.72***	604.2***
	(2.975)	(75.52)
Observations	780	780
R^2	0.277	0.020

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable is the assigned rating in column (1)and the squared difference between the rating and the actual performance outcome in column (2). *Prosocial* is a dummy variable indicating whether the supervisor is classified as prosocial based on the SVO measure. *Signal average* is the average value of the signals observed by the supervisor. Dummies for all treatments (reference group is *NP-NA-S1*) are included. Data from all treatments.

A potential explanation is that, in addition to the weight that subjects place on someone else's payoff, which should lead to more leniency, supervisor's social preferences also affect rating behavior through additional channels. Bieleke et al. (2020), for instance, find that prosocial subjects invest more time and effort to assess the consequences of their actions for others, i.e., they might take the rating task more seriously in particular if bonus

payments hinge on their assessment.[25] Moreover, prosocial subjects may also be more prosocial towards the experimenter and may feel a stronger obligation to provide more accurate ratings. To test this conjecture we run an additional regression of the rating in column (1) of Table 8 where we allow the slope of the rating function to vary with the SVO type. Indeed, prosocials have a significantly steeper slope, i.e., they exhibit less compression and react more to the signal. Their rating function has a lower intercept (at signal = 0) and crosses the rating function of individualistic supervisors at about a signal of 38 which is close to the average performance of 42. Similarly to the effect of having more signals about the agent's performance, the prosocials' rating function is thus steeper.[26] This leads to more rating differentiation while keeping average ratings (and ratings for average signals) more or less equal to those given by individualistic supervisors.[27]

A potential rationale for these findings is that signal processing is cognitively costly and prosocials invest more effort in processing the provided information as they take the assessment task assigned to them more seriously. In turn they may exhibit a stronger willingness to follow the signal. To evaluate that conjecture we can make use of the fact that our experimental software tracked the time supervisors needed for their rating decision: Indeed, while individualistic supervisors submit their rating already after 86 seconds on average, prosocial supervisors spend 110 seconds on the rating page and thus take about 28% more time for the rating task ($p \leq 0.0001$, two-sided t-test).

Finally, we test whether the stronger reliance on the signal and the higher time invested in the evaluation by prosocial supervisors also comes along with an increased rating accuracy. To do so, we replicate the model from column (1) in Table 6 where the squared rating error is the dependent variable and add the supervisor's SVO type as further explanatory variable. The results are shown in in column (2) of Table 8. Indeed, we find that prosocial supervisors also exhibit smaller rating errors compared to individualistic supervisors.

It appears that prosocial supervisors perform the rating task more diligently in particular when their ratings have an impact on agents' pay and thus do not become more lenient than their less prosocial counterparts in this case.[28] The less prosocial supervisors apparently are also more inclined to provide lenient ratings when there is performance

[25] In a similar direction, Grosch and Rau (2017) show experimentally that prosocial subjects are more honest. Chen et al. (2013) show that subjects with higher GPA and SAT scores are more likely to be prosocial according to the SVO, such that prosocials likely have higher cognitive abilities.

[26] In the NP-NA-S1 and the P-A-S1 treatments their slope even tends to be larger than the slope predicted by Bayesian updating.

[27] Note that this effect of prosociality is not predicted by our model (and also not by Golman and Bhatia's (2012) model, see section 4.4), where the slope is only affected by the signal precision. It is also not predicted by larger intrinsic accuracy preferences (γ), which only affect the intercept.

[28] This can be rationalized when prosocial supervisors are also more likely to hold meritocratic fairness norms according to which agents who invest more effort deserve higher bonuses (compare for instance Cappelen et al. (2007), Cappelen et al. (2019)).

pay because leniency comes at no personal costs and thus leniency and lazyness to provide accurate ratings go hand in hand.

Hence, a key insight is that there is a non-trivial relationship between prosocial preferences and rating leniency. Prosocial supervisors are more diligent raters and provide more accurate ratings – and they do so in particular when their ratings matter for the agent's payoffs. This effect counteracts the direct leniency effects of social preferences.

4.7 The Variance of Ratings and the Notion of Rating Compression

A standard result often stressed in the study of subjective performance evaluation is *rating compression* or the so-called *centrality bias*. Prendergast (1999) for instance defines "*Centrality bias refers to a practice where supervisors offer all workers ratings that differ little from a norm.*" (p. 30) or Bol (2011) states that "*centrality bias refers to the tendency of managers to provide ratings that fail to adequately discriminate between subordinates in terms of their respective performance level*" (p. 1559). In field settings, centrality bias is often assessed through the variance of the ratings or by contrasting the variance of ratings with the variance of some observable performance proxy (compare e.g. Bol (2011), Engellandt and Riphahn (2011), Kampkötter and Sliwka (2018)).

In the following we investigate the connection between performance differentiation, which is higher when more signals are available, and rating variance in more detail. To do so, it is instructive to consider again the difference between the treatments in which the supervisor can observe four signals rather than one signal. Indeed our model predicts that the variance in ratings is higher in the S4 treatments as compared to the respective S1 treatments: If we allow for rating mistakes, the variance of observed ratings in our formal framework is

$$V[r] = \left(\frac{\beta + b}{\gamma + \lambda}\right)^2 \sigma_\eta^2 + \frac{\sigma_a^4}{\sigma_a^2 + \frac{\sigma_\varepsilon^2}{n}} + \sigma_\zeta^2 \tag{5}$$

where ζ is a random error term in the rating function with $\zeta \sim N\left(0, \sigma_\zeta^2\right)$. Hence, rating variance is higher when n is larger. The key intuition for this claim is that a higher signal precision implies that supervisors differentiate to a stronger extent between high and low performers and thus ratings vary more.

Figure 5 shows the variance of ratings for each of our experimental treatments with performance pay. Given that the underlying performance distributions are virtually identical the treatments affect the rating variance noticeably. However, and in contrast to the above claim, a higher performance differentiation does not go along with a larger rating variance. In fact, the differences in the variance is not statistically significant (p = 0.496 and p=0.228, F-test of equality of variances) and the variance tends even to be *lower* in the S4 than the S1 treatments. This contrasts the theoretical prediction and also

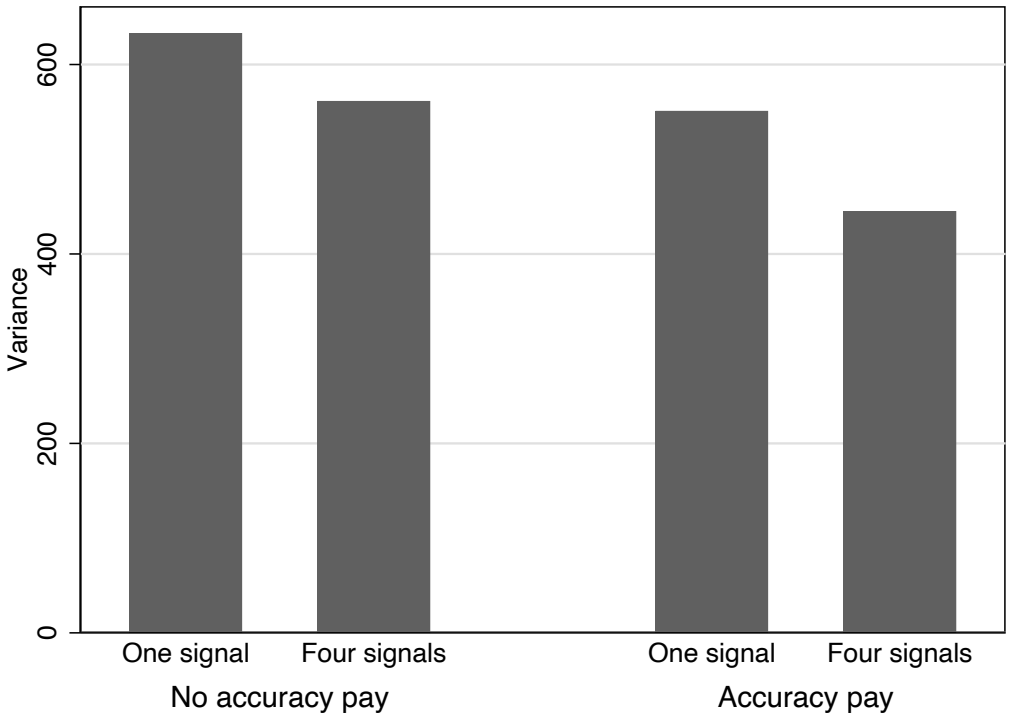


Figure 5: Variance of ratings in treatments with performance pay conditional on accuracy pay and signal precision.

the claim sometimes made in the literature that more accurate ratings are associated with more variation in ratings and less "rating compression".

It is therefore important to explore why this is the case. To do so, it is instructive to consider OLS regressions of the assigned rating on the observed signal average in the four respective treatments to decompose the variance in ratings into the variance explained by the observed signals (the variance of the predicted values) and the remaining variance (i.e. the variance of the residuals) as displayed in Table 9.

Table 9: Decomposing the variance of ratings

	(1) P-NA-S1	(2) P-NA-S4	(3) P-A-S1	(4) P-A-S4
Signal average	5.315***	7.429***	4.631***	6.997***
	(0.993)	(1.148)	(0.839)	(0.932)
Constant	29.74***	19.12***	26.75***	16.01***
	(4.530)	(5.167)	(4.058)	(4.229)
Observations	130	130	130	130
Variance of dependent variable	633.3	561.5	551.1	445.3
Variance of predicted values	115.8	138.4	105.9	136.1
Variance of residuals	517.4	423.2	445.1	309.2
R^2	0.183	0.246	0.192	0.306

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable is the assigned rating. Columns show regressions of the assigned rating on the observed signal average for the performance pay treatments (where signal precision was varied).

As we have seen before and in line with the prediction of the formal model, the slope of the rating function is always larger in the S4 than in the S1 treatments. In turn, the variance of the predicted values is also larger in these treatments. However, the variance of the residuals is larger in the treatments where only one signal is observed.[29] In other words, while indeed due to the more precise information in the S4 treatments ratings vary to a stronger extent with observed signals (which increases rating variance), ratings are also less noisy in this case (which decreases rating variance). It is important to note that this observation is not explained by a model with additive rating mistakes in which the variance in ratings still increases with n (compare (5)). Hence, the results show that the accessibility of more precise rating information reduces the likelihood of mistakes and shifts behavior closer to the prediction implied by rational Bayesian updating. A potential explanation is that due to economizing cognitive costs of information processing supervisors invest more in belief updating when it is more worthwhile to do so, i.e. when more signals are available (compare for instance Kominers et al. (2018) for a formal analysis

[29]This difference in the variance of the residuals is significant in the treatments with accuracy pay but not in the treatments without accuracy pay ($p = 0.040$ and p=0.255 respectively, F-test of equality of variances)

of costly Bayesian inference).

On a more applied level these results suggest some caution in using the variance in performance ratings as a measure of rating compression (or performance differentiation). As our results show, rating variance is driven by both, rating differentiation with respect to performance and rating errors. If more accurate ratings increase (rational) rating differentiation but also reduce rating mistakes, they can go along with a lower rating variance.

4.8 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.[30] 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 (for more details, see Appendix A.1).

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 below their perceived performance.[31]

We find that when workers receive a rating below their performance, their average SVO angle 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"

[30] In contrast to the analysis in section 4.6 where the recipient was another random participant to measure general social preferences, here we use the workers' choices in the SVO task where the supervisor who had assigned the specific rating was the recipient of the respective choice. As we are interested in a behavioral choice rather than a type, we here use the SVO angle as a continuous measure of prosocial choice.

[31] 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.

Table 10: 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.

workers give up 33 cents to increase their supervisor's payoff by 59 cents.

Table 10 reports regressions of the workers' propensity to reciprocate the rating as measured by the SVO angle 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.

5 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 accurate when supervisors determine bonus payments, which supports the notion that there is a tension between different rating purposes as discussed in the introduction. However, rewards for accuracy counteract this effect and reduce leniency 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.

However, we also have found systematic deviations from the model's predictions. For one, supervisors' social preferences do not systematically lead to more generous evaluations. Our analysis has shown that prosocial raters spend more time on the rating and, importantly, their ratings vary more with observed performance which induces a similar effect as the availability of more precise information. This leads to more accurate evaluations counteracting potential leniency effects. Hence, a fear that prosocial supervisors may be tempted to inflate ratings in order to raise subordinate's wages may be unwarranted as prosocial preferences also are associated with a preference for taking the rating task more seriously.

Moreover, we have found that rating variance may not always be an appropriate measure to assess rating differentiation according to performance. Our experimental results highlight that a higher performance differentiation can go along with a lower rating variance. An implication from a more practical perspective is that it may not be a sensible objective for firms to increase rating variance per se, or to consider ratings with a larger variance as superior, but rather to aim at assessing the usefulness of ratings to predict important elements of performance.

References

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.

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., 2011. The determinants and performance effects of managers' performance evaluation biases. The Accounting Review 86, 1549–1575.

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. The 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.

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., 2019. A meritocratic origin of egalitarian behavior. IFN Working Paper, No. 1277 .

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.

Chen, C.C., Chiu, I.M., Smith, J., Yamada, T., 2013. Too smart to be selfish? measures of cognitive ability, social preferences, and consistency. Journal of Economic Behavior & Organization 90, 112–122.

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. The 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.

Engellandt, A., Riphahn, R.T., 2011. Evidence on incentive effects of subjective performance evaluations. ILR Review 64, 241–257.

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

Grabner, I., Künneke, J., Moers, F., 2020. How calibration committees can mitigate performance evaluation bias: An analysis of implicit incentives. The 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.

Kominers, S.D., Mu, X., Peysakhovich, A., 2018. Paying (for) attention: The impact of information processing costs on bayesian inference. Mimeo Harvard University; Available at SSRN 2857978 .

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.

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

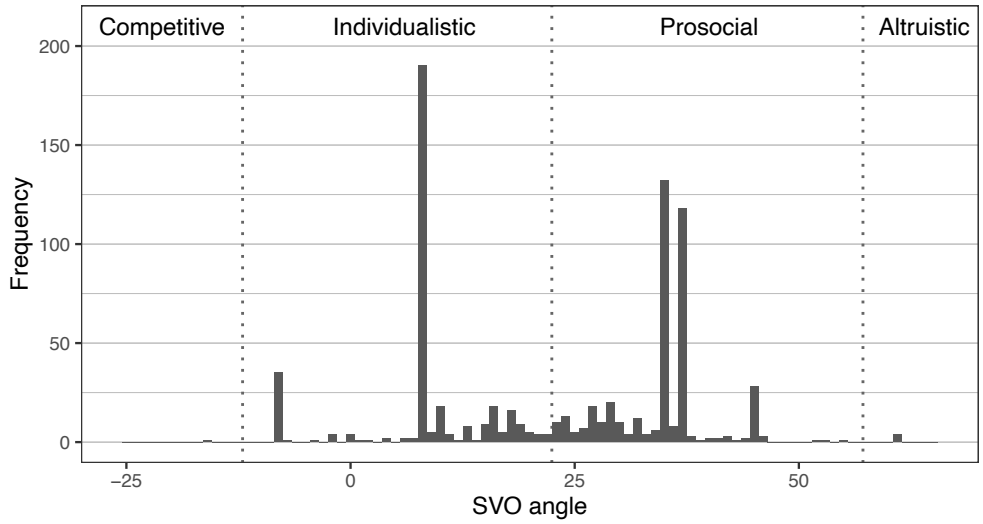
Villeval, M.C., 2020. Feedback policies and peer effects at work, in: Zimmermann, K.F. (Ed.), Handbook of Labor, Human Resources and Population Economics. Springer.

A Appendix

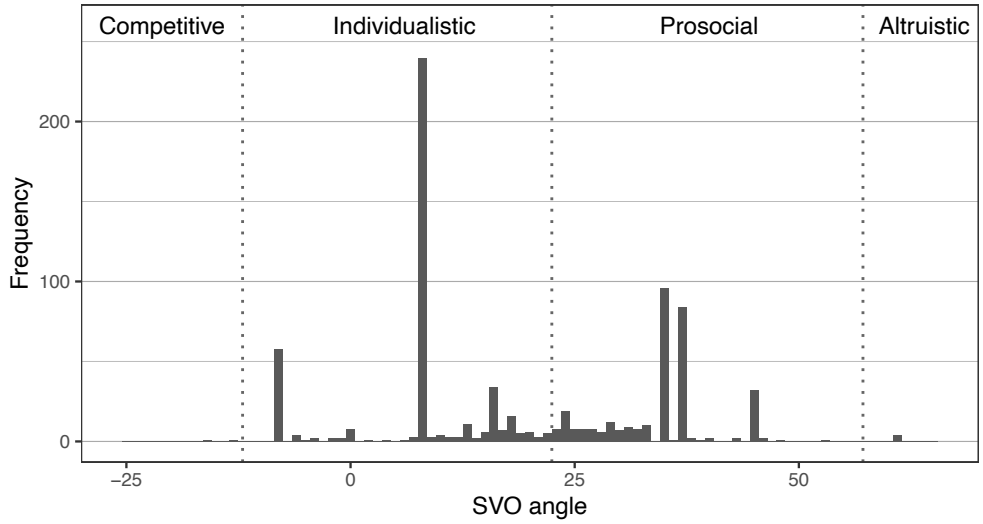
A.1 Social Value Orientation

In this section we describe the SVO classification in greater detail. Our SVO elicitation and the classification is based on the six primary items of the SVO slider task by Murphy et al. (2011). Each item consists of a dictator game where the subject chooses between an amount of money for themselves and an amount of money for another person. For supervisors, this other person is another worker (but not the one they rated), for workers it is the supervisor who rated their work. The SVO angle is computed from the angle between the average money chosen for themselves and chosen for the other (for details, see Murphy et al. (2011)). Four types of social preferences are distinguished in the SVO literature: *altruistic*, where a subject always chooses the option that maximizes the other's payoff, *prosocial*, where a subject either chooses the allocation where the difference between the payoffs is minimized or the allocation where the joint gain is maximized, *individualistic*, where a subject chooses the allocation where their own payoff is maximized, and *competitive*, where the subject chooses the option that maximizes the difference between their own and the other's payoff (with themselves being ahead). Murphy et al. (2011) apply decisions according to these idealized types to the six primary items to determine type boundaries from the resulting SVO angles.

Figure A.1 shows the SVO angle distribution and type boundaries for supervisors and workers in our experiment. A large majority of subjects falls into the individualistic and prosocial categories. There are 4 (4) altruistic and 1 (2) competitive supervisors (workers). Whenever we use the prosocial/altruistic dichotomy in the paper, we add altruistic subjects to the prosocial category and competitive subjects to the individualistic category. Overall, in our experiment there are 45% individualistic and 55% prosocial supervisors and 56.7% individualistic and 43.3% prosocial workers.



(a) Supervisors



(b) Workers

Figure A.1: SVO angle and SVO type classification boundaries

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

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.3 Social Preferences, Rating Leniency & Rating Error

Figure A.2: Rating leniency and rating errors by SVO type