I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16812

# Multi-Rater Performance Evaluations and Incentives

Axel Ockenfels
Dirk Sliwka
Peter Werner

FEBRUARY 2024

DISCUSSION PAPER SERIES

IZA DP No. 16812

# Multi-Rater Performance Evaluations and Incentives

**Axel Ockenfels**
*University of Cologne and Max Planck Institute for Research on Collective Goods*

**Dirk Sliwka**
*University of Cologne, IZA and cesIfo*

**Peter Werner**
*Maastricht University*

FEBRUARY 2024



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

## Multi-Rater Performance Evaluations and Incentives*

We compare evaluations of employee performance by individuals and groups of supervisors, analyzing a formal model and running a laboratory experiment. The model predicts that multi-rater evaluations are more precise than single-rater evaluations if groups rationally aggregate their signals about employee performance. Our controlled laboratory experiment confirms this prediction and finds evidence that this can indeed be attributed to accurate information processing in the group. Moreover, when employee compensation depends on evaluations, multi-rater evaluations tend to be associated with higher performance.

**JEL Classification:** J33, M52

**Keywords:** performance appraisal, calibration panels, group decision-making, real effort, incentives

**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

* All authors gratefully acknowledge financial support of the German Research Foundation (DFG) through the research unit “Design & Behavior” (FOR 1371), Ockenfels and Sliwka gratefully acknowledge funding from the DFG under Germany’s Excellence Strategy – EXC 2126/1– 390838866, and Ockenfels gratefully acknowledges support by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No 741409).

## 1 Introduction and motivation

Performance evaluations are an important task of human resource management as they are the basis for merit pay or promotion decisions. In our study, we develop a theoretical model and a laboratory experiment to compare the performance of single-rater and multi-rater evaluations. Multi-rater evaluations have become a common tool in companies as a part of the formal performance appraisal process. In a survey of large U.S. companies, 54% of the respondents with a formal performance evaluation process reported having a calibration or group review process (Society for Human Resource Management 2011).

In this study, we focus on whether multi-rater performance evaluations improve the accuracy of ratings and whether they in turn increase performance. We build on the framework developed by Prendergast and Topel (1996) to study performance evaluations with information aggregation in groups as proposed by Roux and Sobel (2015). In our model, multiple supervisors receive different noisy signals about the employee's performance and then give a rating that determines the employee's payoff. The model predicts that through simple Bayesian updating, information aggregation in supervisor groups leads to higher evaluation accuracy, and, in turn, to higher performance incentives for the employees.

We then test these predictions in a controlled laboratory experiment that compares a baseline setting in which a supervisor receives a noisy signal on the performance of one employee who works repeatedly on a real effort task with a setting where there are multiple supervisors. In this multi-rater evaluation treatment three supervisors simultaneously receive different signals on one employee's performance. They are made aware that each of them receives different information and know the distribution of the respective noise. Supervisors then discuss the performance of the employee via chat and collectively decide on a rating. The rating process reflects the unanimity rule in that while each supervisor independently fills in an assessment, supervisors only receive a payment if their ratings match.

Consistent with our predictions, we find that multi-rater assessments are associated with more accurate performance ratings compared to the ratings of individual supervisors. This effect is due to the additional information provided by the signals available to other supervisors. That information aggregation plays the crucial role is supported by a further treatment, in which one supervisor receives three signals, and thus individually has access to the same information structure as the three raters in our group treatment. We find the rating accuracy of these individuals is about the same as the accuracy in the treatment with three raters. Both of our treatments with more signals also tend to increase employee performance in later rounds when agents have experienced the higher rating accuracy. Thus, in line with our theoretical predictions, the higher accuracy of ratings based on the aggregation of more information can lead to higher powered incentives driving higher efforts.

Theoretical research on subjective performance evaluations in economics has often been concerned with evaluations made by firm owners, where a key issue is that employers might misrepresent appraisals to save on labor costs (see e.g. Baker et al. 1994, MacLeod 2003).[1] However, most performance evaluations are actually conducted by supervisors who do not have to pay the resulting bonuses out of their own pockets. As a classic literature in psychology (see e.g., Murphy and Cleveland 1995, and Prendergast 1999 for a survey from the economics perspective) has pointed out, supervisors then tend to compress ratings or to be too lenient.[2] Such settings are captured in the Prendergast and Topel (1996) framework, where a supervisor receives a noisy signal about an agent's performance and has to provide a performance rating that trades off a preference for accurate evaluations with favoritism

[1] A firm's commitment to pay a fixed wage sum paired with relative evaluations of employees, such as in tournament systems, can mitigate this problem (Prendergast and Topel 1993, Letina et al. 2020). Moreover, in relational contracts where firms and employees interact on an ongoing basis and firms aim to motivate employees in the future, there is less incentive to negatively distort subjective performance evaluations (see Baker et al. 1994, Lazear and Oyer 2012).

[2] Field data from companies and marketplaces (Moers 2005, Bol 2011, Bolton et al. 2013, Breuer et al. 2013, Ockenfels et al. 2015, Frederiksen et al. 2017, Bolton et al. 2019) confirm that evaluations tend to be too positive ("leniency bias") and to be compressed around some standard ("centrality bias"). While it is not yet fully understood empirically how leniency in evaluations impacts employee performance, several studies suggest that the rating compression reduces performance (Engellandt and Riphahn 2011, Bol 2011, Berger et al. 2012, Kampkötter and Sliwka 2017, Manthei and Sliwka 2019; Kampkötter and Sliwka 2016 provide a survey of subjective performance evaluation practices).

towards the agent. Our model abstracts away from the latter and rather focuses on information aggregation when there are multiple evaluators.[3]

Experimental work on performance appraisals in economics has mostly focused on single supervisor settings. Berger et al. (2012) conduct an experiment in which a supervisor rates multiple employees and find that forced rankings lead to higher performance but also result in more sabotage. Sebald and Walzl (2014) study agents' reciprocal reactions to subjective performance evaluations by a supervisor. Angelovski et al. (2016) find that supervisors tend to be biased in favor of their own hires. Marchegiani et al. (2016) show that ratings that are too lenient are less detrimental to agents' performance than ratings that are too negative. Bellemare and Sebald (2019) find evidence that workers' responses to subjective performance evaluations are related to agents' self-confidence levels. Kusterer and Sliwka (forthcoming) also experimentally investigate the predictive power of the Prendergast and Topel (1996) framework in a single rater setting and find that the model mostly organizes the data but also that supervisors' social preferences are associated with higher rating precision. The only laboratory study we are aware of that considers the aggregation of performance evaluations in groups is Mengel (2021), who investigates how the deliberation process within committees affects subjective evaluations and finds that open deliberation introduces a gender bias in subjective assessments. Unlike our design, this study does not compare single- and multi-supervisor assessments.

In general, group evaluations may have potential advantages over single-rater evaluations, such as the mitigation of distorted evaluations due to favoritism or biased information processing, reducing the risk of collusion between supervisor and employee (as it is more difficult for employees to influence multiple raters than single raters), and improving the coordination of supervisors on the same evaluation standard. Only a few recent studies provide some evidence on multi-rater evaluations based on firm level observational data: Grabner et al. (2020), for instance, find that calibration panels tend to discipline supervisors

[3] Previous extensions of the Prendergast and Topel (1996) model are Golman and Bhatia (2012), who allow for differences in the supervisor's aversion towards favorable and unfavorable errors, Kampkötter and Sliwka (2017), who study bonus dispersion in teams, and Manthei and Sliwka (2019), who investigate performance evaluations when agents work on multiple tasks. The latter two also provide evidence for the respective implications of the model based on field data.

who provide biased information. Demeré et al. (2018) find that calibration committees are associated with lower rating leniency but (surprisingly) with higher rating compression. Bol et al. (2023) compare rating adjustments after calibration rounds and find that posterior ratings are more consistent with the rating distribution desired by the firm. At the same time they observe evidence for strategic behavior of supervisors in the calibration process.[4] Our laboratory experiment provides complementing evidence by studying the causal effects of having multiple raters on information aggregation and induced employee incentives in a setting in which supervisors' interests are aligned.

Our paper also links to research on including raters from multiple layers within the organization (i.e., also colleagues and subordinates in so-called "360-degree" appraisals), which has shown mixed effects. Atkins and Wood (2002) study 360-degree appraisals within a firm, showing that feedback information can be significantly distorted depending on the source of the information. Carpenter et al. (2010) show that peer evaluations lead to workers sabotaging each other under a tournament incentive scheme in the laboratory. Corgnet (2012) observes in a real-effort experiment on teams that equal sharing rules may outperform peer assessments by co-workers. Carpenter et al. (2018) find that peer reports improve performance more strongly under a profit sharing rule than under fixed wages.

Overall, to the best of our knowledge, no previous experimental study has compared single- with multi-rater performance evaluations and their impact on employee performance. Thus, our study can be seen as a starting point for investigating group evaluation processes in more complex settings.

## 2 Theoretical Framework

We build on the framework introduced in Prendergast and Topel (1996) and incorporate multi-rater performance evaluations. A risk averse agent with constant absolute risk

[4] Moreover, based on qualitative interviews with managers from four companies, Lillis et al. (2022) find evidence for the role of calibration in the ranking of employees and in reducing errors and biases in assessments within three of these companies. At the same time, the authors also document potential biases introduced within the calibration process.

aversion $r > 0$ and reservation wage $w_A$ works for a risk neutral principal who makes a take-it-or-leave it contract offer. The agent exerts an effort e at costs $c(e)$ generating a profit contribution $\pi = e + a$, where $a \sim N(m, \sigma_a^2)$ is the agent's ability which is ex-ante unknown to all parties. The agent's wage is given by

$$w = \alpha + \beta \cdot \tilde{\pi},$$

where $\tilde{\pi}$ is a performance assessment. There are $M$ supervisors, $j = 1, \ldots, M$ who provide this assessment. Each supervisor $j$ observes a performance signal $s_j = \pi + \varepsilon_j$ where the $\varepsilon_j \sim N(0, \sigma_\varepsilon^2)$ are independent error terms that capture idiosyncratic views on the agent's performance.[5] Assume that the supervisors have a preference to report an accurate estimate for the employee's performance given their noisy joint information. When the supervisors use their collectively available information each supervisor's expected utility is

$$-E[(\tilde{\pi} - \pi)^2 | s_1, s_2, \ldots, s_M].$$

Our framework thus captures a setting where supervisors have different perceptions on how well the agent performed but are aligned in their view of what constitutes good performance, being aware that their own "subjective" perceptions entail errors of observation. Because differences in supervisors' beliefs are captured by the idiosyncratic signals that they receive, rational supervisors who were to jointly observe all signals thus always agree in their assessment of the agent's performance.

The mean of the observed signals $\bar{s} = \frac{1}{M} \sum_{j=1}^{M} s_j$ is a sufficient statistic for $\pi$, and for given equilibrium efforts $e^*$ the signal mean $\bar{s}$ is normally distributed with prior mean $m + e^*$ and variance

$$V\left[\frac{1}{M}\sum_{j=1}^{M} s_j\right] = V[\pi] + \frac{1}{M^2} V\left[\sum_{j=1}^{M} \varepsilon_j\right] = \sigma_a^2 + \frac{1}{M} \cdot \sigma_\varepsilon^2.$$

[5] One example from practice would be the evaluation processes in a consulting firm where a junior consultant works with different senior consultants who observe the junior consultant's performance in different client projects each. Each of the projects gives a signal that will be affected by the junior consultant's effort and ability but also project specific factors uncorrelated to effort and ability. Hence, each of the senior consultants will bring a different signal on the agent's ability to the table.

The optimal joint evaluation policy $\tilde{\pi}(\bar{s}, M)$ then boils down to computing the least squares estimator of $\pi$ based on the signal mean $\bar{s}$ which is identical to reporting the conditional expectation $E[\pi[\bar{s}]$ such that

$$\tilde{\pi}(\bar{s}, M) = \frac{\sigma_\varepsilon^2(\mathrm{m} + \mathrm{e}^*) + M\sigma_a^2\bar{s}}{M\sigma_a^2 + \sigma_\varepsilon^2}, \quad (1)$$

where the latter follows from applying a standard result on the conditional expectation of normally distributed random variables. Hence, the larger the number of supervisors $M$, the larger is the weight put on the signal mean $\bar{s}$, as this aggregate signal contains more information on the true performance. If, however, $M$ is small, performance evaluations are more compressed towards the prior expectation as information is noisier. This has several implications:

**Proposition 1:** *The larger the number M of supervisors providing signals of performance,*

*(i)* *the stronger do evaluations vary with the signal mean $\bar{s}$, i.e.*

$$\frac{\partial\tilde{\pi}(\bar{s}, M)}{\partial\bar{s}\partial M} > 0,$$

*(ii)* *the higher the efforts exerted by an agent for a given contract, and*

*(iii)* *the larger the principal's expected profits in an optimal contract.*

**Proof:** See Appendix A1.

Hence, when more supervisors evaluate an agent's performance, the evaluations are more closely linked to the agent's actual performance. This leads to higher marginal returns to effort and thus to higher incentives and performance. Finally, multi-rater evaluations also lead to higher profits under optimal contracting: Since the ratings then better reflect true performance differences rather than mere noise, this allows to implement higher powered incentives for risk averse-agents. We will test the first two of these implications in our laboratory experiment.

## 3 Experimental Design

Our experiment builds on the formal framework introduced in Section 2. Each employee in the experiment must work for four minutes per round on a tedious real-effort task that consists of counting 7-digits in randomly generated blocks of digits. A screenshot of the task can be found in the sample instructions in Appendix A4. Upon entering the number of 7-digits in a given block, the employee is presented a new block and proceeds. The employee can pause the working task at any time during each round.[6] A performance evaluation in this setting is an assessment predicting the number of correctly solved blocks by the employee in a particular round. Supervisors cannot directly observe an employee's performance, but rather receive noisy signals about the performance at the end of each round. Each performance signal is determined by the sum of the employee's true performance and a normally distributed error term with a mean of zero and a standard deviation of three blocks.[7] Moreover, together with the performance signal(s), supervisors are presented the distribution of correctly solved blocks in the same round from an earlier experiment in which altogether 40 participants in the roles of employees worked on the task under the same piece rate as in part one.[8] In line with the incentives postulated by our model, we apply a quadratic scoring rule for the supervisor's payoff that links her performance rating to the true performance of the employee. The round payoff for a supervisor is determined as follows:

$$\textit{Payoff supervisor} = \textit{max}\{200 \text{ ECU} - 10 \text{ ECU}*(\textit{rating} - \textit{true performance in blocks})^2, 0\}.$$

Payoffs for supervisors are maximized if their rating matches the true performance of the agent. Yet, if the difference between true performance and estimated performance becomes too large, supervisors' payoffs are fixed at zero to rule out the possibility of losses.[9]

---

[6] The employee can click on a "break"-button on the screen and will be forwarded to a pause screen on which comics are displayed.

[7] This implies that evaluations in our model are not subjective in the sense that supervisors must generate their assessments of performance themselves, which would open up the possibility of further judgment biases which we abstract away from. We rather provide our supervisors with exogenously generated and individually distorted performance signals and then focus on how these signals are processed and aggregated in the evaluation process.

[8] The respective text was: "The distribution of work performance in round X in another experiment had been the following:" followed by a histogram of the performance distribution from the prior experiment. When testing the differences between the distributions from the pilot with the actual distribution in each of our three experimental treatments and separately for each of the 10 rounds with two-sided Mann Whitney U-tests, 29 of the 30 tests are insignificant and one is weakly significant with a p-value of 0.06.

[9] This payoff rule becomes relevant when the rating deviates from the employee's true performance by 5 or more blocks, which was true for only 6% of the supervisor ratings in the experiment (calculated over all treatments).

Our three experiment treatments systematically vary the number of supervisors who rate performance and the number of signals about the employee's performance that the supervisors receive.

**Treatment *1Su1Si*:** In our baseline condition *1Su1Si* (= 1 Supervisor, 1 Signal) one supervisor interacts with one employee in each round and receives one noisy signal about the performance of the employee before she assigns a performance rating.

**Treatment *3Su3Si*:** In treatment *3Su3Si*, three supervisors assess the performance of one employee. Each supervisor receives one individual and private signal about the employee's performance. The noise terms in each of the three signals are independently drawn. The task of the supervisors is then to arrive at a joint performance rating for the employee. To calibrate, supervisors can discuss the rating of the employee via chat for 150 seconds. After the chat, each supervisor individually provides a performance rating. If an agreement is reached and all ratings are identical, the payoffs for each supervisor are determined by the scoring rule described above. However, if at least two performance ratings differ from each other, each supervisor obtains a round payoff of zero[10] and the performance rating is then randomly drawn from the individual ratings.

Treatment *3Su3Si* adds communication between supervisors, plus it changes the performance information available to the supervisors compared to *1Su1Si*, where supervisors base their decision on one performance signal instead of three. To control for the effect of more precise performance information, eliminating the role of group communication, we add treatment *1Su3Si*.

**Treatment *1Su3Si*:** One supervisor interacts with the employee but, in contrast to the control condition, she receives three signals about the employee's performance on which she can base her performance rating. Any differences in evaluation patterns between

[10] The aggregation rule matters, as we discuss in our concluding section, and the performance of different rules will depend on context. For instance, giving all supervisors veto power when favoritism is an issue, will likely complicate negotiations. Because there is no conflict of interest among supervisors in our underlying model, and because other aggregating rules such as majority voting create other strategic issues, we decided to demand unanimity.

treatment *1Su3Si* and treatment *3Su3Si* can then be attributed to the interaction between the supervisors within the calibration panels that goes beyond pure information aggregation.[11]

Supervisors and employees interact with each other for 10 rounds. Our experiment consists of two parts that vary the way how payments for an employee is determined. In the first part (rounds 1 to 5), the employee receives a piece rate for each correctly counted block so that his payoff is determined as

*Payoff employee in rounds 1 to 5 = Number of correct blocks* * 15 ECU.

The goal of the first part of the experiment is to allow supervisors and employees to gain experience with the decision situation, before performance ratings become payoff-relevant for the employee. Note that in this part, ratings already determine payoffs for the supervisors, providing incentives to evaluate accurately. In the second part (rounds 6 to 10), when supervisors are expected to have become experienced with the rating procedure, the payoff for the employee is determined by the rating of the supervisor, allowing us to also study the incentive effects of ratings in our setting. Employees' payoffs in the second part are calculated as follows:

*Payoff employee in rounds 6 to 10 = Rating supervisor* * 15 ECU.

The piece rate per block remains the same as in the first part of the experiment, but the payment is no longer determined by the true performance, but rather by the estimate of the supervisor. As a result, distortions in ratings directly affect payments.[12]

Supervisors and employees are matched with each other for the entire 10 rounds of the experiment (partner matching). In every round, supervisors rate the employee, subject to the treatment variations described above. Supervisor learn their signals but not the true

[11] Evidence from several experimental studies suggests that groups may perform better because of the social interaction *per se* (Charness and Sutter 2012 survey the literature). In addition, some evidence suggests that groups might under some circumstances exhibit less socially oriented behavior. Taken together, these findings would suggest for our setting that groups of supervisors may be influenced to a lesser extent by social concerns towards the employee than individual supervisors.

[12] If supervisors submit different ratings in *3Su3Si*, one of the performance ratings is randomly picked for the employee's payoff.

performance during the experiment. Only at the end of the laboratory session are they informed about the number of blocks the employee solved correctly in each of the rounds. On the other hand, employees get to know their true performance after each round and thus can judge the accuracy of their supervisor(s). We note that while supervisors do not know the full distribution of employee performance the information about the distribution of performance in the pilot experiment should align priors. We will return to this issue when we discuss our results.

Prior to the start of the experiment, participants were assigned the role of either an employee or a supervisor. Supervisors and employees were seated in different rooms to minimize social interaction. After the 10 rounds of the main experiment, we collected measures for inequity aversion (Dannenberg et al. 2007), cognitive reflection (Frederick 2005), intelligence (with Raven matrices) and some basic demographic information about the experimental participants.[13] We conducted altogether 10 experimental sessions at the Cologne Laboratory for Economic Research from October 2016 to April 2017. Participants were recruited via the online recruitment system ORSEE (Greiner 2015). The experiment code was created with the software z-tree (Fischbacher 2007). We collected data for altogether 200 subjects in our experiment.[14] Average payments accounted for 26.76 Euro (standard deviation 4.52 Euro) including a show-up fee of 4 Euro for sessions that lasted between 1.5 to 2 hours. Due to the partners matching in our experiment, we collected altogether 22, 24 and 27 statistically independent observations for treatments *1Su1Si*, *1Su3Si* and *3Su3Si,* respectively.

## 4 Results

[13] We find no sizeable differences in demographics across treatments. We note, however, that the gender composition is weakly significantly different ($p = 0.052$, Chi-Square test; see Table A1 in the Appendix).
[14] In most sessions, the experiment was conducted on two computer servers simultaneously. In one session, one of the two servers stopped working so that the experiment had to be stopped for some participants. For our analysis, we exclude these additional 8 subjects (treatment *3Su3Si*).

Table 1 displays descriptive statistics of the employees' performance and the performance ratings assigned by the supervisors in the three treatments across the two parts of the experiment.

**Table 1:** Means (Std. dev.) of performance and ratings

| | Performance | | Rating | |
|---|---|---|---|---|
| Treatment | Part 1 | Part 2 | Part 1 | Part 2 |
| *1Su1Si* | 8.51 (3.29) | 8.50 (3.50) | 8.34 (2.88) | 8.73 (2.76) |
| *1Su3Si* | 7.47 (2.35) | 8.66 (2.71) | 7.52 (2.73) | 8.54 (2.62) |
| *3Su3Si* | 7.53 (3.05) | 8.15 (3.85) | 7.67 (2.30) | 8.40 (3.69) |

Note - The table lists the average number of correctly counted blocks and the average ratings per round, separately for each experimental treatment and part. Standard deviations are listed in parentheses.

The first thing to note is that in line with the model average ratings match average performance (in number of correctly counted blocks per round) quite well. Moreover, in our baseline treatment *1Su1Si* performance and ratings stay relatively stable across the two parts (round 1-5 versus round 6-10), whereas performance tends to increase in the second part where ratings become payoff relevant for the employee in the treatments with multiple signals/raters. In the following we test the hypotheses implied by the formal model in more detail.

## *4.1 Evaluations*

Our first key hypothesis is that the increase in the number of evaluators/signals shifts the sensitivity of the rating to the signal: By equation (1), the optimal evaluation by supervisors $\frac{\sigma_\varepsilon^2(m+e^*)}{M\sigma_a^2+\sigma_\varepsilon^2}+\frac{M\sigma_a^2}{M\sigma_a^2+\sigma_\varepsilon^2}\bar{s}$ is linear in the respective signal average $\bar{s}$. So, we estimate this equation in linear random effects regressions. By part (*i*) of our proposition, the reported performance evaluation is predicted to more strongly depend on observed signals if more signals are observed (as $\frac{3\sigma_a^2}{3\sigma_a^2+\sigma_\varepsilon^2}>\frac{\sigma_a^2}{\sigma_a^2+\sigma_\varepsilon^2}$). On the other hand, the regression intercept is predicted to be smaller when there are more signals as $\frac{\sigma_\varepsilon^2(m+e^*)}{3\sigma_a^2+\sigma_\varepsilon^2}<\frac{\sigma_\varepsilon^2(m+e^*)}{\sigma_a^2+\sigma_\varepsilon^2}$. Also, if supervisors follow Bayes' rule and there are no further group interaction effects, we predict that there are no differences in evaluations between treatments *1Su3Si* and *3Su3Si*.

To make quantitative predictions for the evaluations, we substitute the expected equilibrium performance $m + e^*$ in Equation (1) with the mean true performance (8.112), and $\sigma_a$ with the standard deviation of the true performance (3.196). By our experiment design, $\sigma_\varepsilon = 3$. The resulting predictions for the constants and slope coefficients for the three treatments are reported in Table 2.

**Table 2:** Rational evaluations as a function of performance signal

| | Treatment *1Su1Si* | Treatment *1Su3Si* | Treatment *3Su3Si* |
|---|---|---|---|
| Predicted slope | 0.53 | 0.77 | 0.77 |
| Predicted constant | 3.80 | 1.84 | 1.84 |

To test the predictions, we estimate the following simple specification separately for each of the three treatments:

$$rating_{it} = \alpha + \beta * signal_{it} + \epsilon_{it},$$

where $rating_{it}$ is the rating of subject $i$ in period $t$, and $signal_{it}$ is the (aggregated) signal observed by the supervisors in the respective treatment (i.e. the average of the signals in treatments *1Su3Si* and *3Su3Si*). The results reported in Table 3 confirm the first part of our proposition: The signal is positively and highly significantly correlated with the performance ratings that employees received in all specifications. In line with our prediction, we find that the coefficient size for the signal is larger in treatments *1Su3Si* and *3Su3Si* than in the control condition. Hence, supervisors indeed react more sensitively to the signals here. At the same time, the size of the coefficients is similar for treatments *1Su3Si* and *3Su3Si,* showing that the group interaction during the calibration process in treatment *3Su3Si* does not further increase the sensitivity of supervisors to the performance signals.

**Table 3:** Signals and supervisor decisions

| Dependent Variable | (1)<br>Rating<br>*1Su1Si* | (2)<br>Rating<br>*1Su3Si* | (3)<br>Rating<br>*3Su3Si* | (4)<br>Rating All |
|---|---|---|---|---|
| Signal | 0.469*** | 0.708*** | 0.648*** | 0.471*** |
| | [0.076] | [0.053] | [0.065] | [0.076] |
| *1Su3Si* | | | | -2.144*** |
| | | | | [0.715] |
| Signal × *1Su3Si* | | | | 0.241*** |
| | | | | [0.094] |
| *3Su3Si* | | | | -1.331* |
| | | | | [0.785] |
| Signal × *3Su3Si* | | | | 0.178* |
| | | | | [0.101] |
| Constant | 4.356*** | 2.231*** | 3.013*** | 4.334*** |
| | [0.589] | [0.414] | [0.528] | [0.583] |
| Observations | 220 | 240 | 270 | 730 |
| $Chi^2$-value | 37.94 | 178.1 | 97.97 | 314.4 |

Note - Standard errors clustered on the level of supervisors in *1Su1Si* and *1Su3Si* (on the level of supervisor groups in *3Su3Si*) are given in brackets. *** p<0.01, * p<0.1. The table reports the results of linear regression models with random effects on the level of supervisors (supervisor groups in *3Su3Si*). The variable “Signal” refers to the performance signal in treatment *1Su1Si* and the average of the three performance signals in treatments *1Su3Si* and *3Su3Si*.

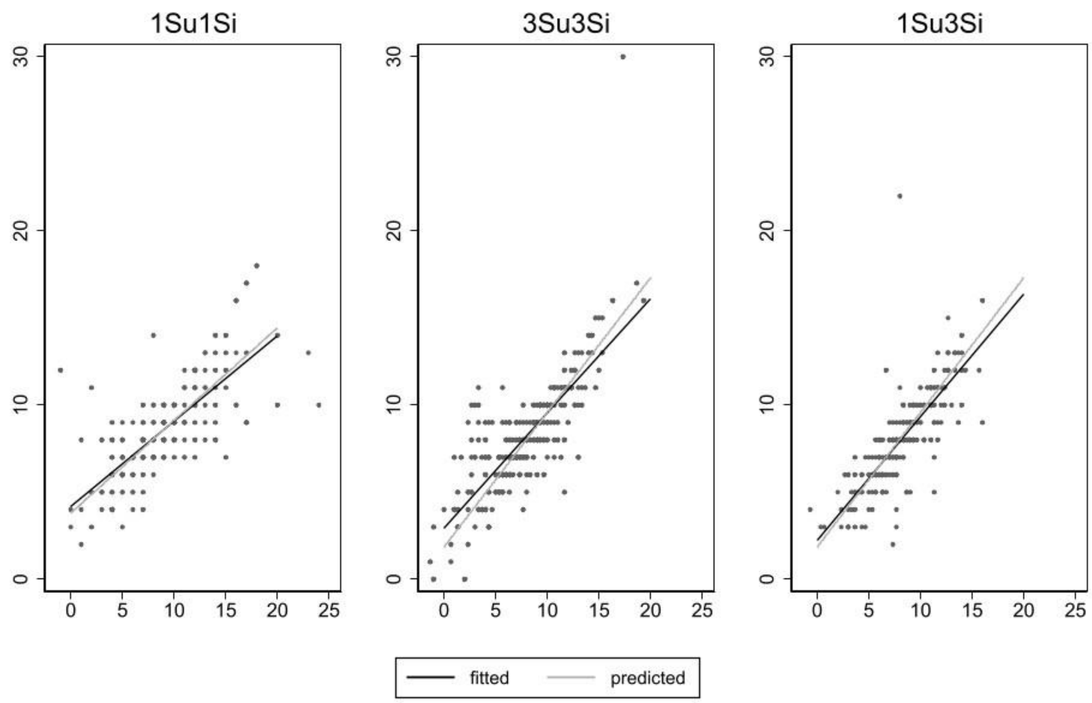
A similar conclusion is reached from the model in column (4) focusing on the interaction terms of the treatment dummies with the signal. Both interaction variables Signal x *1Su3Si* and Signal x *3Su3Si* are positive and (marginally) significant, showing a stronger correlation between the signal and the performance rating in these treatments.[15] At the same time, the interaction terms are not significantly different from each other ($p$ = 0.457, two-sided Wald test).[16] Moreover, in line with the predictions, the coefficients of *1Su3Si* and *3Su3Si* are negative and (marginally) significant in Model 4, indicating smaller constants in these treatments relative to the baseline condition.

[15] In a single-rater setting Kusterer and Sliwka (forthcoming) also find that more accurate signals lead to a stronger signal sensitivity of ratings.

[16] If we estimate Model 4 reported in Table 3 separately for the first and the second half of the experiment (see Table A2 in the Appendix), we find a stronger response of supervisors to signals in the treatments mainly for the second half. Hence, it appears that supervisors (groups) require some time to learn how to properly use the performance signals in our setting.

Figure 1 shows scatterplots for the observed (aggregated) signals and evaluations for the three treatments separately. The grey line shows the pattern predicted by Bayes' rule (i.e., using the respective intercept and slope reported in Table 2) and the black line shows the fitted OLS estimates. The data appear to be mostly well organized by the predictions implied by rational Bayesian updating, especially in the treatments with one supervisor. The fit is somewhat weaker in the three-supervisor treatment but the qualitative treatment differences appear to be well in line with the key predictions of the model.

**Figure 1:** Predicted and Fitted Performance Evaluations




The x-axis of each graph in Figure 1 refers to observed (average) signals; the y-axis of each graph refers to ratings.

All in all, the experimental results appear to be broadly consistent with the Bayesian rational model under the assumption of common priors among supervisors.

## 4.2 Evaluation Accuracy

A further prediction of our model is that ratings in *1Su3Si* and *3Su3Si* are more accurate. Table 4 lists the mean deviation of the evaluation from the employee's true performance (measured in absolute values) separately for the three treatments and the two parts of the experiment.

**Table 4:** Average supervisor accuracy per treatment

| Treatment | Deviation rating from true performance (in number of blocks, absolute values) | |
|---|---|---|
| | Part 1 | Part 2 |
| *1Su1Si* | 2.12 | 2.37 |
| *1Su3Si* | 1.73 | 1.52 |
| *3Su3Si* | 1.94 | 1.75 |

Our model suggests that the precision of ratings is higher in treatment *3Su3Si* than in *1Su1Si*. We find that although the between-treatments difference in absolute deviations between ratings and the employee's true performance is not significant for the first part of our experiment ($p = 0.960$), it is strong and statistically significant for the second part ($p = 0.013$, two-sided MWU test): The calibration process within the group leads to more accurate performance ratings. Also, as predicted, treatment *1Su3Si* achieves a similar degree of accuracy as the group evaluation treatment *3Su3Si* in both parts of the experiment ($p = 0.214$ for part 1 and $p = 0.634$ for part 2, respectively, two-sided MWU tests).[17] This supports the view that the superiority of the group evaluation in our setting can be attributed to the additional performance signals which make it easier for the supervisors to arrive at an appropriate evaluation, but that the communication and interaction process within the group *per se* does not have a sizeable impact. Moreover, we find virtually no conflict between the supervisors in treatment *3Su3Si*: Calculated over all rounds of the experiment, supervisor teams reached an agreement (i.e. the same performance rating) in 269 out of 270 cases, mirroring the strong incentives to arrive at a joint rating in our setting.

[17] Mirroring the differences between *3Su3Si* and the control condition, comparing treatments *1Su3Si* and *1Su1Si* yields significant differences for part 2 ($p = 0.003$), but not for part 1 ($p = 0.223$).

Table 5 below reports the results of simple regressions of the deviation between the rating and the true performance on treatment dummies for each part (columns (1) and (2)). These regressions indicate that there are learning effects as the availability of more signals significantly increases the accuracy in the second but not the first part of the experiment. Moreover, as in the previous descriptive analyses, we do not find differences between treatments *1Su3Si* and *3Su3Si* in the models (the respective two-sided Wald tests yield significance levels above 0.1).[18]

**Table 5**: Deviation between rating and true performance

| | 1 | 2 |
|---|---|---|
| Dependent variable | Deviation Part 1 | Deviation Part 2 |
| Signal | 0.050 | 0.006 |
| | [0.047] | [0.033] |
| *1Su3Si* | -0.310 | -0.856*** |
| | [0.235] | [0.283] |
| *3Su3Si* | -0.106 | -0.619** |
| | [0.231] | [0.312] |
| Constant | 1.667*** | 2.316*** |
| | [0.344] | [0.340] |
| Observations | 365 | 365 |
| $Chi^2$-value | 2.73 | 9.37 |

Note - Standard errors clustered on the level of supervisors in *1Su1Si* and *1Su3Si* (on the level of supervisor groups in *3Su3Si*) are given in brackets. *** p<0.01, ** p<0.05, * p<0.1. The reference category in the models is treatment *1Su1Si*. Model 1 (2) refers to the first (second) part of the experiment. The table reports the results of linear regression models with random effects on the level of supervisors (supervisor groups in *3Su3Si*). The variable “Signal” refers to the performance signal in treatment *1Su1Si* and the average of the three performance signals in treatments *1Su3Si* and *3Su3Si*.

## *4.3 Impact on Performance*

Claim (*ii*) of our proposition predicts that performance is higher when it is assessed by more supervisors or when more performance signals are available: When more signals are

[18] As Bayesian updating requires cognitive capabilities, we also explore the potential role of supervisors' cognitive abilities for the evaluations. We do so by integrating the supervisors' score in the Cognitive Reflection test (CRT, Frederick 2005) elicited after the main part of the experiment in the model (Table A3 in the Appendix lists the results). In treatments *1Su1Si* and *1Su3Si*, the CRT score refers to the score of the individual supervisor; in treatment *3Su3Si* it stands for the average CRT score of the three supervisors per matching group. In both parts, supervisor (groups) with higher CRT scores achieve a higher accuracy.

available, supervisors should put more weight on the (aggregate) signals, which in turn leads to steeper incentives as performance ratings depend to a stronger extent on observed signals. And indeed, as we have seen in Figure 1 and Table 2, the corresponding slopes are higher in the experiment. As a result, marginal returns to effort are higher in treatments *1Su3Si* and *3Su3Si* as compared to *1Su1Si*. If employees correctly anticipate this, or learn about the relationships between effort and compensation 'on the job', work efforts should be higher in the second part of the experiment (periods 6 to 10) in which the payoff for the employee is determined by the supervisors' evaluation. To test this, we regress employee performance in part 2 on treatment dummies and prior performance. We use the performance in number of correctly solved blocks in a given round as the dependent variable. We control for the employee's prior performance using the average performance in blocks per round in the first part of the experiment where employees work under piece rate incentives and thus are not affected by the evaluations of the supervisors. To study whether agents learn over time that incentives are steeper in the treatments where more signals are available we also investigate the treatment effects only in the final two rounds of the experiment. Table 6 presents the respective regression results.

We find in both specifications that performance is significantly higher in treatment *1Su3Si* where one principal receives three signals. The point estimate for the treatment *3Su3Si* is also positive but much smaller in magnitude and not significantly different from performance in *1Su1Si* in Model 1 that includes all observations from the second part. Model 2 from Table 6 that includes only observations from rounds 9 and 10 supports the view that employees needed time to learn about the improved accuracy in *3Su3Si*. In this model, the dummy variable *3Su3Si* is larger and marginally significant, indicating higher performance in this treatment relative to the control condition at the end of the experiment.[19] Hence, employees seem to initially underestimate the reliability of group evaluations, yet gradually learn about higher evaluation accuracy as the labor relationship progresses. The observation that multiple signals induce stronger performance incentives for the employees

[19] Comparing the estimated coefficients for the treatment dummies with two-sided Wald tests does yield significant differences ($p = 0.102$ for Model 1 and $p = 0.729$ for Model 2).

particularly towards the end of the experiment (Model 2), when subjects may be more tired or less concentrated, tends to strengthen our conclusion.

**Table 6:** Effects on employee performance

| Dependent Variable | 1<br>Performance<br>Part 2 | 2<br>Performance<br>Part 2, last rounds |
|---|---|---|
| *1Su3Si* | 1.197** | 1.221** |
| | [0.478] | [0.606] |
| *3Su3Si* | 0.628 | 1.064* |
| | [0.489] | [0.583] |
| Avg. performance (Part 1) | 0.996*** | 0.981*** |
| | [0.115] | [0.100] |
| Constant | 0.021 | 0.019 |
| | [0.786] | [0.818] |
| Sample | Part 2 | Rounds 9 and 10 |
| Observations | 365 | 146 |
| R-squared | 0.519 | 0.499 |

Note - Standard errors clustered on the level of experimental employees are given in brackets. *** p<0.01, ** p<0.05, * p<0.1. The reference category in the models is treatment *1Su1Si*. The table reports the results of OLS regressions. The variable "Average performance (Part 1)" refers to the average round performance per employee calculated over the first five periods of the experiment.

## 5 Conclusion

Our model predicts, and our laboratory experiment confirms that collective evaluations by multiple raters can be more accurate than assessments by a single rater. This improvement is a result of the aggregation of scattered information from different supervisors. In addition, our model and experiment show that multi-signal performance ratings tend to positively affect employee performance, as the higher rating accuracy strengthens the generated incentives – although we note that this effect appears only in later rounds, when agents have experienced the higher rating accuracy.

Our model and laboratory setting can be extended to investigate the role of several complexities in the evaluation process that may be of additional importance in real-world settings. For example, it is straightforward to investigate the role of the number of supervisors involved in the group evaluation: The principal's profits are concave in the number of evaluators in our model, because the marginal informational value of each further signal is decreasing. Thus, when there is a fixed cost for each individual evaluation, one can compute an optimal number of evaluators. There may also be other „diseconomies of scale“ due to cost of communication and coordination arising in larger groups that may be explored through extensions of our theoretical and laboratory models.

The scope of the framework could also easily be extended by varying the difficulty of evaluating performance. In the language of our model, the difficulty of judging performance is captured by the variance of the idiosyncratic error terms. From this perspective, group evaluations would become more beneficial for tasks that are more difficult to evaluate, as the marginal information value of additional signals increases.

Our setting minimizes the scope for biases in the rating process such as employee favoritism or stereotyping of particular demographic groups, as all interactions are anonymous. Supervisors who are inclined to bias evaluations in order to promote certain employees can substantially complicate information aggregation, both in theory and in the laboratory.[20] We hypothesize that under such conditions, group communication will play an even more crucial role in mitigating bias, adding a further benefit to the informational advantage demonstrated in our study. Generally, groups tend to make fewer mistakes than individuals (Charness and Sutter 2012), and norm-setting, confronting biased colleagues, and moderator intervention are all potentially useful group mechanisms in such scenarios (e.g., Johnson and Johnson 2009, Kahneman et al. 2021).

Another potential extension is to allow for one supervisor to have access to more precise information than the others, such as a line manager who observes her employee more closely than others. A simple way to incorporate this into our framework is to assume that one supervisor observes a more precise signal or, equivalently, multiple signals of a given precision. If the joint objective is to maximize accuracy, this does not change the basic mechanics of the model. This would lead the group to give more weight to that supervisor’s information, and this should improve information aggregation when there is no conflict of

[20] For instance, when one supervisor is biased and the others are aware of this bias, they may want to counter the bias by distorting their ratings in the other direction.

interest. However, supervisors who work more closely with the agent being evaluated may also have closer social ties, which could lead to favoritism. Then there is a trade-off between the better information of this supervisor and potential bias due to favoritism. In such cases, it may be preferable to disregard the input of supervisors with vested interests, e.g., by using majority voting rather than unanimity, by excluding supervisors with a conflict of interest from group discussions, or by establishing a more hierarchical communication structure with a group moderator who has the power to weight the input of group members. Other potentially interesting factors may include supervisors' personalities, which may influence the frequency and quality of their contributions to the deliberative process, with, for instance, more extroverted supervisors potentially having a stronger influence on evaluations.

We are confident that such research, including under the controlled conditions of the economic laboratory, will help to complement field studies in ways that will prove useful in designing institutions for more successful and more accurate performance evaluations.

**References**


Angelovski, Andrej, Jordi Brandts, and Carles Sola. 2016. Hiring and escalation bias in subjective performance evaluations: A laboratory experiment. *Journal of Economic Behavior & Organization* 121:114-29.

Atkins, Paul W.B., and Robert E. Wood. 2002. Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology* 55 (4):871-904.

Baker, George, Robert Gibbons, and Kevin J. Murphy. 1994. Subjective performance measures in optimal incentive contracts. *Quarterly Journal of Economics* 109 (4):1125-56.

Bellemare, Charles, and Alexander Sebald. 2019. Self-confidence and reactions to subjective performance evaluations. IZA Discussion Paper No. 12215.

Berger, Johannes, Christine Harbring, and Dirk Sliwka. 2012. Performance appraisals and the impact of forced distribution—an experimental investigation. *Management*

*Science* 59 (1):54-68.

Bol, Jasmijn C. 2011. The determinants and performance effects of managers' performance evaluation biases. *Accounting Review* 86 (5):1549-75.

Bol, Jasmijn C., Andson Braga de Aguiar, and Jeremy Lill. 2023. Agency conflict within calibration committees. Working Paper.

Bolton, Gary, Ben Greiner, and Axel Ockenfels. 2013. Engineering trust: reciprocity in the production of reputation information. *Management Science* 59 (2):265-85.

Bolton, Gary E., David J. Kusterer, and Johannes Mans. 2019. Inflated reputations: uncertainty, leniency, and moral wiggle room in trader feedback systems. *Management Science* 65 (11): 5371-91.

Breuer, Kathrin, Petra Nieken, and Dirk Sliwka. 2013. Social ties and subjective performance evaluations: an empirical investigation. *Review of Managerial Science* 7 (2):141-57.

Carpenter, Jeffrey, Peter Hans Matthews, and John Schirm. 2010. Tournaments and office politics: Evidence from a real effort experiment. *American Economic Review* 100 (1):504-17.

Carpenter, Jeffrey, Andrea Robbett, and Prottoy A. Akbar. 2018. Profit sharing and peer reporting. *Management Science* 64 (9):4261-76.

Charness, Gary, and Matthias Sutter. 2012. Groups make better self-interested decisions. *Journal of Economic Perspectives* 26 (3):157-76.

Corgnet, Brice. 2012. Peer evaluations and team performance: when friends do worse than strangers. *Economic Inquiry* 50 (1):171-81.

Dannenberg, Astrid, Thomas Riechmann, Bodo Sturm, and Carsten Vogt. 2007. Inequity aversion and individual behavior in public good games: an experimental investigation. ZEW Discussion Papers No. 07-034.

Demeré, B. William, Karen L. Sedatole, and Alexander Woods. 2018. The role of calibration committees in subjective performance evaluation systems. *Management Science* 65 (4):1562-85.

Engellandt, Axel, and Regina T. Riphahn. 2011. Evidence on incentive effects of subjective performance evaluations. *ILR Review* 64 (2):241-57.

Fischbacher, Urs. 2007. z-Tree: Zurich toolbox for ready-made economic experiments.

*Experimental Economics* 10 (2):171-78.

Frederick, Shane. 2005. Cognitive reflection and decision making. *Journal of Economic Perspectives* 19 (4):25-42.

Frederiksen, Anders, Fabian Lange and Ben Kriechel. 2017. Subjective performance evaluations and employee careers. *Journal of Economic Behavior & Organization* 134:408-29.

Golman, Russell, and Sudeep Bhatia. 2012. Performance evaluation inflation and compression. *Accounting, Organizations and Society* 37 (8):534-43.

Grabner, Isabella, Judith Künneke, and Frank Moers. 2020. How calibration committees can mitigate performance evaluation bias: an analysis of implicit incentives. *Accounting Review* 95 (6):213-33.

Greiner, Ben. 2015. Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association* 1:114-125.

Johnson, David W., and Roger T. Johnson. 2009. An educational psychology success story: social interdependence theory and cooperative learning. *Educational Researcher* 38 (5): 365-79.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. Noise: a flaw in human judgment. New York: Hachette Book Group.

Kampkötter, Patrick, and Dirk Sliwka. 2016. The complementary use of experiments and field data to evaluate management practices: The case of subjective performance evaluations. *Journal of Institutional and Theoretical Economics* 172 (2):364-89.

Kampkötter, Patrick, and Dirk Sliwka. 2017. More dispersion, higher bonuses? On differentiation in subjective performance evaluations. *Journal of Labor Economics* 36:511-49.

Kusterer, David, and Dirk Sliwka. Forthcoming. Social preferences and the informativeness of subjective performance evaluations. *Management Science*.

Lazear, Edward, and Paul Oyer. 2012. Personnel economics. In *Handbook of Organizational Economics*, ed. Robert Gibbons and John Roberts. Princeton University Press.

Letina, Igor, Shuo Liu, and Nick Netzer. 2020. Delegating performance evaluation. *Theoretical Economics* 15:477-509.

Lillis, Anne M., Mary A. Malina, and Julia Mundy. 2022. The role of subjectivity in mitigating incentive contracting risks. *Accounting Review* 97 (1):365-88.

MacLeod, W. Bentley. 2003. Optimal contracting with subjective evaluation. *American Economic Review* 93 (1):216-40.

Manthei, Katrin, and Dirk Sliwka. 2019. Multitasking and subjective performance evaluations: Theory and evidence from a field experiment in a bank. *Management Science* 65 (12):5861-83.

Marchegiani, Lucia, Tommaso Reggiani, and Matteo Rizzolli. 2016. Loss averse agents and lenient supervisors in performance appraisal. *Journal of Economic Behavior & Organization* 131:183-97.

Mengel, Friederike. 2021. Gender bias in opinion aggregation. *International Economic Review* 62:1055-80.

Moers, Frank. 2005. Discretion and bias in performance evaluation: the impact of diversity and subjectivity. *Accounting, Organizations and Society* 30:67-80.

Murphy, Kevin R., and Jeanette N. Cleveland. 1995. Understanding performance appraisal: Social, organizational, and goal-based perspectives. Thousand Oaks: Sage.

Ockenfels, Axel, Dirk Sliwka, and Peter Werner. 2015. Bonus payments and reference point violations. *Management Science* 61 (7):1496-1513.

Prendergast, Canice. 1999. The provision of incentives in firms. *Journal of Economic Literature* 37:7-63.

Prendergast, Canice, and Robert Topel. 1993. Discretion and bias in performance evaluation. *European Economic Review* 37:355-65.

Prendergast, Canice, and Robert Topel. 1996. Favoritism in organizations. *Journal of Political Economy* 104:958-78.

Roux, Nicolas, and Joel Sobel. 2015. Group polarization in a model of information aggregation. *American Economic Journal: Microeconomics* 7:202-32.

Sebald, Alexander, and Markus Walzl. 2014. Subjective performance evaluations and reciprocity in principal–agent relations. *Scandinavian Journal of Economics* 116:570-90.

Society for Human Resource Management. 2011. SHRM Poll: Performance management and other workplace practices.