



Make a difference? – An experiment on the effectiveness of performance rating policies

Christine Harbring^a, Stella Simons^{a*}

^a RWTH Aachen University, 52062 Aachen, Templergraben 64, Germany

ARTICLE INFO

JEL classification:

C91

M5

Keywords:

Performance appraisals

Rating policies

Performance

Centrality bias

Leniency bias

Laboratory experiment

ABSTRACT

Performance appraisals by supervisors are often used in organizations as a basis for bonus and promotion decisions. These subjective appraisals are often prone to supervisors' biases (e.g., centrality and leniency bias) resulting in ratings that seem not sufficiently differentiated. Previous research shows that forcing supervisors to differentiate in subjective performance appraisals by implementing a forced distribution system may increase worker productivity. However, forced distribution systems pose a variety of other problems. In a laboratory experiment, we investigate two alternative rating policies which aim to foster differentiation and thus productivity. Firstly, supervisors are simply advised to differentiate across different appraisal levels. We find that this simple recommendation results in less lenient and less compressed performance ratings and tends to increase workers' performance. Secondly, we analyze how supervisors evaluate workers' performance when the rewards, which are tied to ratings, may not exceed a certain budget. We detect that specifying a financial budget mitigates the leniency bias and increases workers' performance. Under both rating policies, the performance differences wear off over the course of the experiment.

1. Introduction

Supervisors' performance appraisals are used in companies for several purposes such as salary administration or promotion decisions (Cleveland et al., 1989). Therefore, from an employers' perspective, performance appraisals are an important aspect of decision-making. However, many positions often lack objective performance measures (Grund & Przemeck, 2012), such that performance appraisals are often based on subjective evaluations.¹ A wide range of empirical literature shows that subjective performance appraisals may suffer from biases (Bol, 2011; Jawahar & Williams, 1997; Kneeland, 1929; Moers, 2005). Two common biases are the leniency and centrality bias (Landy & Farr, 1980; Moers, 2005; Prendergast, 1999). Assuming that the distribution of actual performance levels of employees is somewhat in line with a normal distribution, supervisors' evaluations should also resemble a normal distribution. In the case of a leniency bias, the average rating of supervisors deviates upwards from the average performance, i.e., evaluations awarded are too good (Prendergast, 1999). In the case of a centrality bias, supervisors do not use the entire range of possible ratings, i.e., they hardly differentiate between employees (Kampkötter & Sliwka, 2016).

The leniency and centrality bias often occur at the same time because rating scales almost always have an upper limit which implies that lenient ratings are simultaneously compressed ratings (Berger et al., 2013). From an economic perspective, both biases are expected to lower the performance incentives when employees' rewards are tied to ratings (Berger et al., 2013). Due to such compressed ratings, the marginal return to effort decreases and the incentives to exert effort are reduced. Consequently, employees may lower their productivity (Bol, 2011).

* Corresponding author.

E-mail addresses: christine.harbring@org.rwth-aachen.de (C. Harbring), stella.simons@org.rwth-aachen.de (S. Simons).

¹ The terms "appraisal", "rating", "evaluation" and "grade" are used interchangeably throughout the paper.

We investigate the effectiveness of two rating policies which are supposed to mitigate at least one of the above mentioned rating biases and which are frequently encountered in organizational reality. First, we analyze how the recommendation to differentiate across different appraisal levels affects supervisors' performance appraisals and workers' subsequent performance. The recommendation to give evaluations according to a certain distribution is a common practice in companies and easy to implement. According to the Linked Personnel Panel (LPP),² 13 percent of the companies that use performance appraisals recommended to give evaluations in line with a predefined distribution in 2014 (Kampkötter & Sliwka, 2016). We test whether such recommendations effectively reduce biases in performance appraisals.

Second, we explore how supervisors assign performance appraisals when the rewards, which are tied to ratings, may not exceed a certain budget. This second rating policy is related to the concept of bonus pools that are commonly used to allocate incentive payments (Benson & Sajjadi, 2018). Companies often have a limited amount of money, i.e., a budget, that is disposable for bonus payments. We analyze whether financial budgets help to reduce leniency bias in performance appraisals and improve workers' performance. In addition, we analyze not only the performance appraisal decisions of supervisors, but also the subsequent performance of workers. Under both rating policies, we examine how supervisors' performance appraisals impact workers' performance.

Implementing a forced distribution rating system (FDS) is one obvious other approach to overcome the insufficient differentiation in performance appraisals. An FDS requires supervisors to adhere to a predetermined distribution of performance appraisals, so that only a certain fraction of employees can receive the best grades. Inevitably, an FDS leads to differentiated performance appraisals and therefore prevents lenient and compressed ratings. In a laboratory experiment, Berger et al. (2013) showed that forcing supervisors to differentiate in the subjective performance evaluations by implementing an FDS may increase worker productivity. However, FDS may also foster a competitive work environment, as employees compete for a given number of good grades, potentially increasing productivity but also encouraging counterproductive behaviors, such as sabotaging coworkers (Berger et al., 2013; Harbring & Irlenbusch, 2011). Thus, an FDS may not be the most efficient and desired option to decrease rating biases and enhance performance.

In our laboratory experiment, we test two rating policies that serve as alternatives to FDS. First, we investigate whether the recommendation to differentiate appraisals reduces rating biases and increases performance compared to the baseline treatment. Second, we assess the effect of a financial budget, which will lead to less lenient performance appraisals, assuming that supervisors assign mostly good ratings in the baseline treatment.

The experiment can be summed up as follows. Each experimental group consisted of one supervisor and three workers who had to work on a real-effort task for several periods. After each period, the supervisor received feedback on the workers' performance that was then rated on a five-point scale by the supervisor determining the workers' payoff. The supervisor's payoff was based on the workers' performance. The experiment was based on three treatments: (i) In baseline treatment *Base*, supervisors were unrestricted in their evaluation decisions (ii) In the treatment *Rec*, supervisors were given the recommendation to differentiate across appraisal levels (iii) In the treatment *Bud*, supervisors were not allowed to exceed a certain financial budget while assigning the evaluations associated with monetary bonuses. We explore supervisors' ratings and workers' subsequent performance under the rating policies compared to the baseline treatment without such rating policies. Our laboratory data enable a detailed analysis of the treatment effects. We consider multiple measures of rating behavior (grades, grade distributions, span of grades) and performance (individual and group output), allowing us to gain a nuanced understanding of both supervisors' and workers' behavior.

Interestingly, it turns out that the introduction of a recommendation is effective, i.e., we find less lenient and less compressed performance ratings, and workers' performance levels tend to initially increase compared to the baseline treatment. However, performance differences disappear over the course of the experiment. Secondly, we find that specifying a financial budget mitigates the leniency bias and increases workers' performance as well. Again, the performance differences fade out over the course of the experiment. While both policies seem successful first in enhancing performance levels, the underlying drivers of these increases, however, seem to be quite different in both treatments. Higher differentiation in ratings improves performance when supervisors are recommended to differentiate, but not when their ratings are restricted by a financial budget. The performance increase in the budget treatment may instead reflect fairness considerations or positively reciprocal reactions of workers to more equitable, compressed ratings.

The next section gives an overview of the related literature. Section three features the description of the experimental design, and section four develops the hypotheses that are investigated in this study. The experimental results are presented in section five and are discussed in section six. Section seven concludes.

2. Related Literature

As outlined in the introduction, we aim to investigate whether recommending differentiated performance appraisals will mitigate the leniency and centrality bias and consequently lead to a higher performance. There is already some laboratory and field evidence that suggests a positive relationship between higher degrees of differentiation and a higher performance. First, we start with presenting the field studies.

Kampkötter and Sliwka (2018) use a panel data set from the German banking and financial service industry to show that a higher degree of differentiation in performance ratings is effective in increasing performance, at least for individuals at the middle or high level of the hierarchy. Also Bol (2011) takes field data to demonstrate that an increasing centrality bias negatively affects performance improvement. Likewise, the results from Engellandt and Riphahn (2011) indicate that differentiated performance appraisals lead to a higher subsequent performance. By using performance evaluation data from Korean governmental companies, Ahn et al. (2008) report

² The LPP is a German dataset that links employer and employee data.

that there is a significant, positive relationship between the degree of differentiation of appraisals and companies' performance improvement.³ While these field studies have the advantage of a high external validity by utilizing field data from organizations, they rely on proxies for performance measures such as individual bonus payments or overtime work (Engellandt & Riphahn, 2011; Kampkötter & Sliwka, 2018). Our experiment has the decisive advantage that we measure individual performance exactly and objectively. We introduce a recommendation for evaluations and a budget while keeping everything else constant. Thus, in this randomized controlled study, we can clearly investigate how supervisors' ratings affect workers' performance at an individual level.

In addition to the above mentioned field studies, there are also laboratory experiments that report a positive relationship between differentiated appraisals and a higher subsequent performance. Berger et al. (2013) examine how an FDS - and herewith, a forced differentiation of performance appraisals - affects workers' performance. Our study extends the work of Berger et al. (2013) by following their experimental design and integrating their baseline data into our analysis. We add two additional treatments - *Bud* and *Rec* - which were conducted within the same experimental series as the treatments in Berger et al. (2013).⁴

Berger et al. (2013) demonstrate that an FDS leads to a higher performance compared to a baseline setting where supervisors are not restricted in their evaluations. Another main result is that an FDS harms performance if the workers have the possibility to sabotage their co-workers. In the FDS setting, workers significantly increase their sabotage activities due to the competitive situation that is created by the FDS.⁵

Whereas Berger et al. (2013) investigate the impact of an FDS on workers' performance, we investigate the impact of the recommendation to differentiate ratings and the impact of a financial budget on workers' performance. Unlike an FDS, the recommendation does not force supervisors to allocate differentiated ratings, but they can freely decide whether to follow it. As the recommendation is non-binding, classical economic theory would not predict an effect. However, the recommendation reframes the rating context compared to *Base*, thus functioning as a nudge towards more differentiated ratings. The recommendation merely introduces the idea of differentiated ratings to the supervisors, offering a more realistic and less disruptive alternative than introducing an FDS.

Similar to an FDS, the financial budget restricts supervisors' decision-making by limiting their possibility to allocate lenient ratings due to insufficient financial resources. In organizations, the budget for bonuses is restricted quite often; the *Bud* treatment, is motivated by bringing the setting closer to reality. In *Bud*, supervisors can give differentiated ratings; however, unlike an FDS, they can still opt to rate workers equally at mediocre or bad rating levels. Consequently, the *Bud* treatment models a realistic restriction and allows us to investigate how differentiated ratings impact performance when equal ratings are still possible but resources are restricted. To conclude, in the paper by Berger et al. (2013) two extreme forms of settings are compared: (i) Allowing supervisors to autonomously decide on ratings (ii) Restricting supervisors' decision range to a specific distribution of ratings. The treatments in the current paper model two settings which may be somewhere in between these extremes and are close to organizational reality as supervisors are often trained and advised to follow a certain distribution of grades, and bonus budgets are typically restricted.

In a different laboratory experiment, Loberg et al. (2021) investigate how an FDS affects performance when employees are either working alone or in a team. In the setting where employees work alone, an FDS leads to a higher individual performance. However, in a team-based setting, an FDS leads to a decreasing performance. Moreover, Loberg et al. (2021) find that supervisors tend to distribute the bonus pool in a centralized manner when they can freely decide how to distribute the bonus pool.

There is also a strand of literature that examines how differentiated appraisals influence the willingness to exert effort instead of the actual performance. Since the willingness to exert effort is closely related to and often the requirement for actual performance, we also draw on results from Mursita and Nahartyo (2022) and Trapp and Trapp (2019). In a real-effort laboratory experiment, Mursita and Nahartyo (2022) show that the willingness to exert effort is significantly higher in a situation without centrality bias compared to a situation with centrality bias. Likewise, Trapp and Trapp (2019) demonstrate that the willingness to exert effort is significantly lower when centrality bias is present in a vignette study. This effect is mainly driven by above-average performers.

Generally, all of the described studies suggest a positive relationship between differentiated performance appraisals and employee performance, but some draw attention to restrictions regarding this relationship. Kampkötter and Sliwka (2018) restrict this positive relationship to employees at middle and high hierarchy levels. Moreover, Berger et al. (2013) indicate that the FDS only increases performance when there is not the possibility to sabotage the co-workers. Loberg et al. (2021) demonstrate that an FDS is counterproductive if employees work in a team, and only a team performance measure exists. Hence, situational circumstances should be taken into consideration when differentiated performance appraisals, and especially those that are enforced by an FDS, are applied. Although there are some factors that should be considered, former empirical evidence strongly suggests a positive relationship between differentiated performance appraisals and a higher performance. Given the positive performance effects of more differentiation in appraisals, the question arises how organizations can promote differentiation without the drawbacks of FDS. Therefore, we aim to investigate whether differentiated appraisals can be triggered via a simple recommendation to differentiate.

3. Experimental Design

In our laboratory experiment, all subjects were divided into groups of four people consisting of one supervisor and three workers.

³ This study differs from the previously mentioned studies because here, the performance of companies is evaluated and not directly the performance of employees. Since companies are run by people that strongly impact an organization's performance, this study also gives an indication of how people react to compressed appraisals.

⁴ The *Base* treatment was conducted in May 2009, *Bud* in February 2010, and *Rec* in March 2010.

⁵ In our study, we do not investigate the effect of sabotage.

The workers had to work on a real-effort task where they had to count how often the number seven is contained in blocks of random numbers.⁶ The experiment consisted of two parts, and in each part the workers had to work on this simple cognitive task for eight periods (à 2.5 minutes).⁷ For each correctly counted block each worker received two points and for each incorrect block the worker lost half a point. Therefore, the output for each period was calculated by subtracting the points of the incorrect blocks from the points of the correct blocks. Instead of counting the sevens, the workers had the possibility to push a timeout button. When this button was pushed, the worker's own screen was blocked for 20 seconds and she received 25 cents. This option was implemented to model opportunity costs of working, participants typically have outside of the lab.⁸ After each period, the workers were informed about their own performance (output, number of correct and incorrect blocks) and their co-workers' performance.

The supervisors' task was to evaluate the workers' performance after each period. They had to rate the workers' performance on a scale from one to five (1 = best; 5 = worst). After each period, the workers were privately informed about their own rating, but not about the co-workers' ratings. Depending on their rating, the workers received a bonus payment (see Table 1). With a decreasing grade, the bonus payments decreased in 2.5 € steps. Accordingly, a worker's payoff is entirely determined by the bonus payment and any payment generated by pressing the timeout button. The supervisor's payoff depended on the workers' performance and was determined by multiplying the output of the three workers by 30 cents. At the end of the experiment, one period of each part was randomly selected to be payoff relevant.

The experiment was run with three different treatments which varied regarding their rating policies (see Table 2 for an overview). In *Baseline (Base)*, supervisors could assign ratings as they wanted. The treatment, *Recommendation (Rec)*, differed by additionally giving the recommendation to adhere to a certain rating distribution. In *Budget (Bud)*, supervisors were limited by a 15 € budget per period. The recommendation and the budget were common knowledge, i.e., also the workers were informed about them in the instructions.⁹

All subjects were ranked according to their preliminary-period output. The top four subjects were matched in a group. Then, the next best four subjects were assigned to a group and so on: until the worst four players formed a group. The top performer of each group was appointed supervisor. Individuals with homogenous abilities were matched in one group to avoid obvious ratings because of differences in the subjects' abilities. This matching procedure reflected typical workplace scenarios where comparisons are typically only made among employees of similar abilities who are comparable. Subjects knew that they retained their assigned roles and remained within the same groups. However, at no point in time subjects were informed about the group matching procedure to avoid strategic considerations. As a result, they could neither manipulate their effort in the preliminary period to affect group placement nor knew that their peers had similar abilities. The idea was to get as close to the difficulties of situations with performance appraisals in organizations as possible.

The experiment, conducted in 2010 as part of a larger project on performance appraisals, was programmed by using the software "z-Tree" (Fischbacher, 2007), and subjects were recruited via ORSEE (Greiner, 2004). The 192 participants were students from the University of Cologne with just over half of the sample being students of business or economics. Each treatment involved 64 subjects (16 independent groups), consisting of 48 workers and 16 supervisors. Each treatment was conducted over two sessions, with 32 subjects (8 groups) per session. The mean age of the whole sample was 23.55 years with women accounting for 44.27% and men for 55.73%.

In the following, the procedure of each session is described. At the beginning of a session, all subjects received instructions on the sample task and the preliminary period (see Figure 1 for an overview of the procedure). The sample task consisted of one block of random numbers in which the number of sevens had to be determined. There was no time constraint for the sample task, and the subjects could only precede when the correct answer was given. The sample task should ensure that every subject understood the real-effort task that had to be accomplished in the preliminary period, in part one, and in part two of the experiment.

After the sample task, all subjects participated in the preliminary period. Here, subjects had to work on the real-effort task for 2.5 minutes, and they received a piece rate of 10 cents for their output. In the preliminary period, the subjects had the possibility to push a timeout button which blocked their screen for 20 seconds, and they received 8 cents for each time the timeout button was pushed. The preliminary period was implemented to measure the subjects' innate ability in performing the real-effort task.¹⁰ After the preliminary period, the instructions of part one were distributed. Before the start of the first part, the subjects answered several test questions to ensure that they understood the group composition and payoff structure. The instructions of part two were distributed before the second part of the experiment. No changes were made between the first and second part, i.e., participants were informed that they retained their original role and group assignments. This two-part structure allowed us to test the effect of refreshing the recommendation which may have diminished in influence over the course of the first part. The recommendation was initially presented in the instructions before the first part and repeated in the instructions prior to the second part. To ensure comparability with the *Rec* and *Base* treatments, *Bud* was also designed with the two-part structure.

At the end of a session, a few decision games were played to measure subjects' social preferences. The games, adapted from Dannenberg et al. (2007) and originally developed by Blanco et al. (2011), enable the classification of supervisors into four types: selfish, altruistic,

⁶ Visual search tasks are well-established real-effort tasks for measuring individual performance in laboratory experiments (Abeler et al., 2011; Mohnen et al., 2008).

⁷ While workers are working on the real-effort task, supervisors do not have any assigned task.

⁸ By implementing the timeout button, we increase workers' costs of working on the real-effort task. This option allows workers to decide for something different than the task, providing them with some positive utility similar to a work arrangement where they can choose to interrupt their work and do something they like. Incentivizing alternatives to working allows the experimenter to get additional insights into participants' behavior.

⁹ Although workers are not informed about their co-workers' grades, they may infer this information in *Bud*, given their awareness of the limited budget. As our treatment is supposed to resemble budget restrictions in organizations, it is an intentional element of the treatment design that indirect information about co-workers' grades may become known in *Bud*.

¹⁰ The timeout button in the preliminary period is also simulating opportunity costs. By ability, we do not only refer to the task performance, but also include the willingness to perform if there is an incentive not to perform.

Table 1
Bonus payments workers.

Rating	Bonus Payment
1	10.00 €
2	7.50 €
3	5.00 €
4	2.50 €
5	0.00 €

Table 2
Overview treatments.

Treatment	Treatment abbreviation	Description
Baseline	<i>Base</i>	Supervisor unrestricted in rating decision
Recommendation	<i>Rec</i>	Recommendation to differentiate: → Supervisor was recommended to give one worker grade one or two, another worker grade three and the third worker grade four or five
Budget	<i>Bud</i>	Bonus payments were not allowed to exceed budget (15 €)

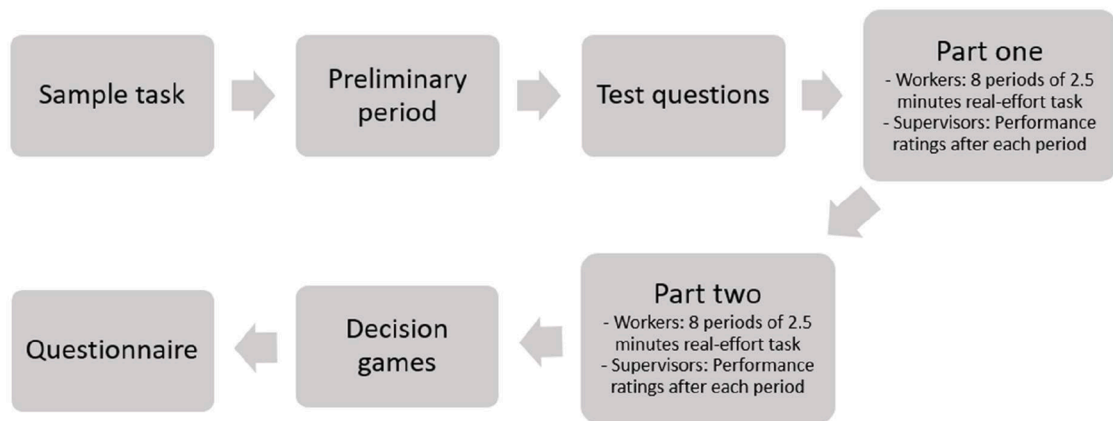


Figure 1. Procedure.

envious and equity oriented.¹¹ Additionally, all subjects had to answer a questionnaire which was divided into a specific part that was different for the supervisors and workers, and a general part that was answered by both. The questionnaire started with various statements regarding the supervisors' or workers' motives. The statements had to be rated for the first and second part of the experiment separately. Then, some more general questions followed such as the Big Five Inventory (Rammstedt & John, 2007), items regarding positive and negative reciprocity (Dohmen et al., 2009) and socio demographic information such as gender or age.

4. Hypotheses

In this section, we present our expectations regarding supervisors' rating behavior and workers' performance. As mentioned before, the supervisors received the recommendation to differentiate across different appraisal levels in *Rec*. They were told to always give one worker a good grade; another worker should receive a mediocre grade; and the third worker should be rated with a bad grade. The recommendation can be conceived as nudging (or a framing effect) since we expect it to affect the supervisors' decision-making process leaving the monetary incentives untouched.

There are also some empirical studies investigating the effect of recommendations on subjects' behavior in different experimental games (Brandts & Macleod, 1995; Bruns & Perino, 2021; Croson & Marks, 2001). Bruns and Perino (2021) have reported that a recommendation increases the propensity to provide the recommended contribution to the public good. Also Croson and Marks (2001) show that in a situation, where each individual values the public good differently, the recommendation enhances the probability of recommended contributions. Accordingly, there is some evidence that individuals consider recommendations in their decisions. In our experiment, the recommendation may trigger the supervisors' consideration of the utility of differentiated appraisals. The supervisors may start thinking about the advantages or disadvantages of a differentiation and may better understand the possible incentive effect, and therefore, act in line with the recommendation. Since the recommendation is exogenously given, and the workers know about the recommendation, it may serve as an excuse for

¹¹ See Berger et al., (2013) for a description of the variable classification. We are interested in supervisors' social preferences, as prior research suggests that evaluation behavior may be influenced by the rater's personality traits (Bernardin et al., 2000; Bernardin et al., 2016).

the supervisors to also give unpleasant ratings. This may also encourage the supervisors to apply differentiated appraisals. Furthermore, the recommendation may establish a norm (Kahneman & Miller, 1986), and thus, supervisors may allocate the ratings accordingly.

Following previous studies, we expect lenient and compressed ratings in the baseline treatment (Bol, 2011; Jawahar & Williams, 1997; Kneeland, 1929; Landy & Farr, 1980; Moers, 2005). In contrast to this baseline treatment, we expect more differentiated performance appraisals when implementing the recommendation to differentiate in *Rec*. In addition, the ratings are presumed to be less lenient because the recommendation suggests to assign not only good grades, but also mediocre and bad grades.

Hypothesis 1a. The recommendation to differentiate (*Rec*) leads to a higher differentiation in performance appraisals compared to the baseline treatment (*Base*) with no recommendation.

Hypothesis 1b. The recommendation to differentiate (*Rec*) leads to less lenient performance appraisals compared to the baseline treatment (*Base*) with no recommendation.

In our experiment, lenient and compressed ratings are closely related: Given the upper limit of our rating scale, lenient ratings inherently lead to compression. Thus, the recommendation to give differentiated ratings aims not only to enhance differentiated ratings, but also to mitigate lenient ratings. In the budget treatment, ratings are inevitably less lenient assuming that ratings are lenient and compressed in our baseline treatment.

The experiment was designed so that workers' payoffs directly depend on the ratings they receive by the supervisor: Higher ratings result in a higher payoff. Under the second rating policy, supervisors have a limited budget for allocating the payoffs (ratings). Therefore, it is not feasible to give all workers good grades because supervisors lack sufficient financial resources. Assuming that supervisors assign lenient ratings in the baseline treatment, ratings in the budget treatment *Bud* should be less lenient due to the lack of financial resources.

While a financial budget will inevitably prevent lenient ratings, it will not necessarily lead to more differentiated performance appraisals. A centrality bias is still feasible for a given budget constraint, as supervisors could still allocate similar mediocre or bad grades. If the budget is set too strictly, it will even enforce compressed ratings because the supervisors are not able to differentiate anymore. Accordingly, we carefully set a budget that prevents lenient ratings, but still offers sufficient scope for differentiated appraisals.

Although supervisors still have the option to allocate differentiated appraisals, we expect them to actually differentiate less. The reason is that supervisors are confronted with a trade-off: Whenever they intend to assign a good grade, they are forced to assign a worse grade to another worker. Thus, a good grade is associated with negative external effects for the other workers. Though, based on previous findings,¹² we assume that participants have an aversion against giving bad ratings. As a result, supervisors may avoid assigning bad grades by assigning more or less equal ratings to workers. Therefore, we hypothesize that the budget leads to less differentiated performance appraisals.

Hypothesis 1c. Setting a financial budget (*Bud*) leads to less differentiated performance appraisals compared to the baseline treatment (*Base*) with no budget restrictions.

From an economic perspective, the leniency and centrality bias are expected to lower the performance incentive when the payoff is tied to ratings (Berger et al., 2013). According to agency theory, receiving a payoff which depends on the individual performance increases the motivation to exert effort (Eisenhardt, 1989; Prendergast, 1999). But, if a higher performance does not translate into a higher payoff, because the leniency or centrality bias lead to a mismatch between the actual performance and the payoff (Bol, 2011), the pay-for-performance incentive may be disrupted.

As hypothesized before, we expect that the recommendation leads to a higher differentiation in performance appraisals compared to the baseline treatment. Under differentiated performance appraisals, ratings align with actual performance.¹³ In contrast, compressed ratings result in similar ratings,¹⁴ even if performance levels vary. This mismatch between actual performance and ratings alters incentive structures: Top performers would have to perform at a higher level to receive the same rating as without the centrality bias, while low performers would need to increase performance considerably to achieve a better rating because the rating is already inflated. Hence, compressed ratings lead to a decreasing marginal return to effort reducing the incentive to perform.

If supervisors follow the recommendation to differentiate, the recommendation will also help to reduce lenient ratings because supervisors abstain from allocating similarly good grades to all workers. Thus, a high differentiation in performance appraisals also depicts less lenient ratings. In the case of lenient ratings, not just top performers may receive good grades, but potentially also middle performers and even low performers. Therefore, workers can perform less to achieve the same rating as in a situation with less or no lenient ratings (Bol, 2011). Here, the performance threshold to obtain a good rating is lower compared to stricter performance appraisals, and thus, the incentive to exert effort is reduced.

Hypothesis 2a. We assume that the recommendation to differentiate (*Rec*) results in greater differentiation in performance appraisals which leads to a performance increase.

Hypothesis 2b. We assume that the financial budget (*Bud*) results in less differentiation in performance appraisals which leads to a

¹² That individuals tend to assign lenient ratings is investigated for example by Moers (2005), Landy and Farr (1980) or Bol (2011).

¹³ Top performers receive good grades, middle performers receive mediocre grades, and low performers are evaluated with bad grades. Top, middle and low performers are identified by ranking the workers' output in their group in each period. Hence, the worker with the highest output in a group is the top performer.

¹⁴ Top, middle, and low performers may rather obtain similar or even the same grades.

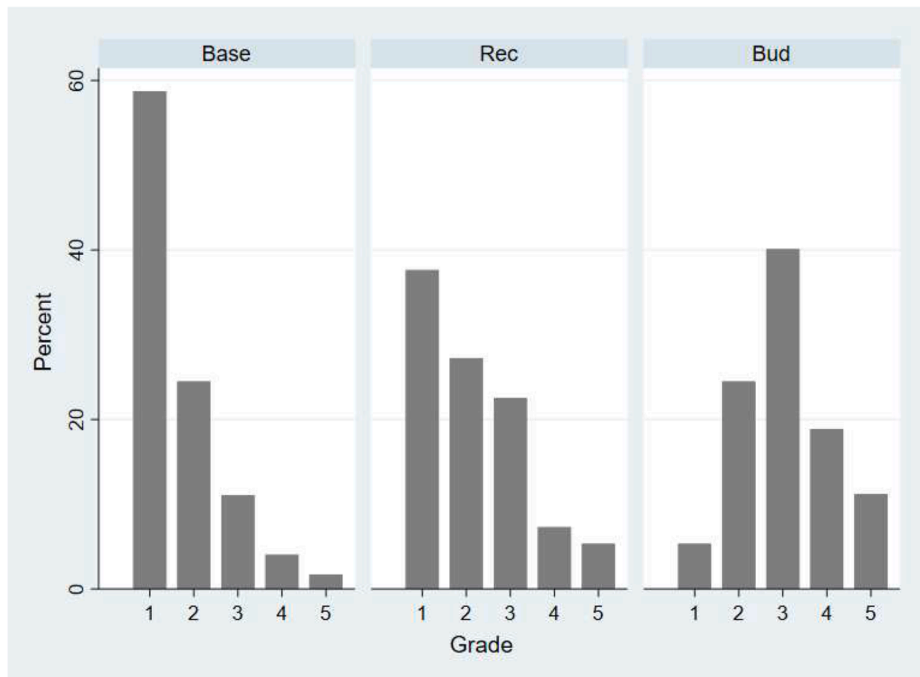


Figure 2. Grades Assigned by Supervisors Periods 1-16¹⁵.

performance decrease.

Our assumptions regarding the mechanisms through which the recommendation to differentiate and the budget constraint influence supervisors' ratings are not explicitly tested. However, the post-experimental questionnaire provides valuable insights into supervisors' rating decisions. As proposed in *Hypotheses 2a* and *2b*, we expect that both rating policies affect workers' performance, and we explicitly test whether differentiated ratings serve as a mechanism driving higher performance. Additional insights into mechanisms are provided by the questionnaire and summarized in subsections 5.1.5 and 5.2.3.

5. Results

In the following, we present the experimental results. First, we consider supervisors' rating behavior in more detail, and second, we investigate how the different rating policies impact workers' performance.

5.1. Supervisors' Rating Behavior

First, we focus on the grades and the grade distributions that are assigned by the supervisors. Then, we analyze whether the supervisors consider relative performance standards while rating the workers. Next, the degree of differentiation in performance appraisals is featured in more detail. Finally, we provide more insights into supervisors' behavioral drivers by analyzing the post-experimental questionnaire.

5.1.1. Grades Assigned by Supervisors

As depicted by Figure 2, almost no worker receives a bad grade (grade "4" or "5") in *Base* while almost 60 % of the workers receive the best grade "1". The average grade in *Base* is 1.7 (see Table A1). This result implies that the supervisors hardly use the whole range of ratings and rather assign pleasant (lenient) ratings when they can freely decide about performance appraisals.

In *Rec*, the average grade of 2.2 is numerically higher than in *Base* (1.7), indicating worse grades. In contrast to *Base*, the best grade "1" is assigned less frequently (38 %) whereas bad grades are allocated slightly more often. Applying the Mann-Whitney U test (MWU test),¹⁶ the ratings in *Rec* are significantly different from the ratings in *Base* ($z = -2.49$, $p\text{-value} = 0.01$).¹⁷ Hence, in *Rec* the ratings

¹⁵ Since the number of observations is central to the analysis, we briefly outline the data structure. For workers, each treatment includes 768 individual-level observations based on 48 workers observed over 16 periods. For supervisors, each treatment includes 256 observations based on 16 supervisors over 16 periods. At the group level, there are a 256 observations per treatment based on 16 groups observed over 16 periods. In Figure 2, all observations are reported at individual worker level.

¹⁶ Here, the MWU test is applied because two independent samples are compared. Throughout all analyses in this paper, the two-sided MWU test is applied.

¹⁷ The average grade of workers across all 16 periods in each group is calculated, resulting in an average value for each group. Each group is statistically independent from other groups as interaction only takes place within the group. These group averages are compared between *Base* and *Rec/Bud*.

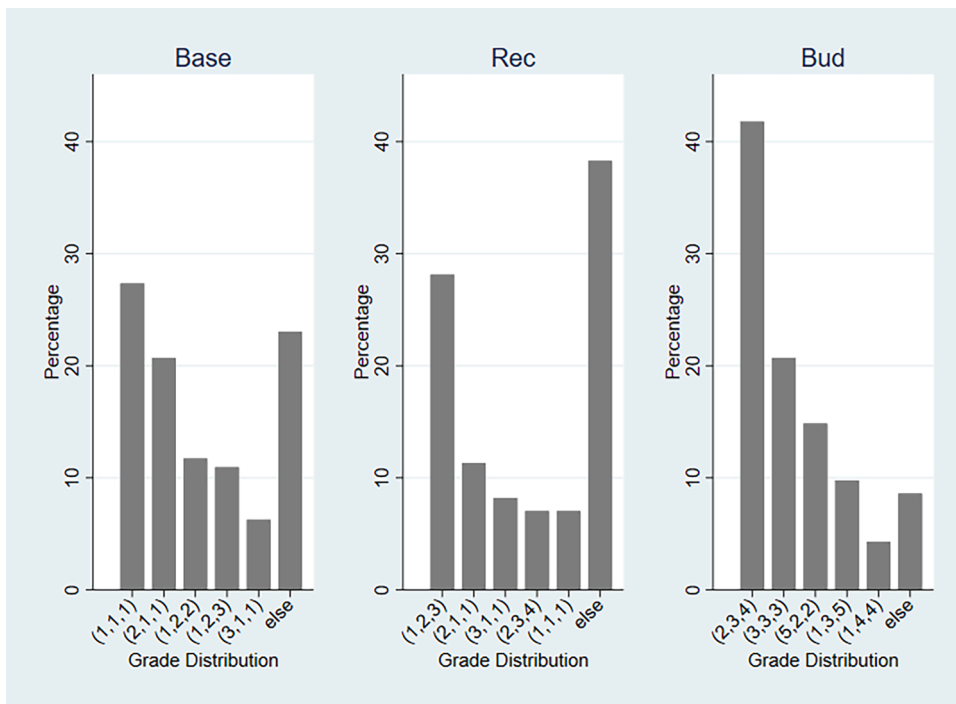


Figure 3. Percentage of Grade Distributions Periods 1-16²⁰.

seem less lenient and more differentiated compared to *Base* which is in line with *Hypothesis 1a* and *Hypothesis 1b*. Also in *Bud*, the ratings differ significantly from *Base* ($z = -4.85$, $p\text{-value} = 0.00$). In *Bud*, the average grade (3.0) is numerically higher than in *Base* (1.7), indicating worse grades. It can be observed that grade “1” is received by only 5 % of workers in *Bud*, making it less frequent compared to *Base*. Instead, grade “3” is most frequently selected (40 %), and the supervisors distribute the less favorable grades more frequently in *Bud*. We conclude that the ratings in *Bud* are less lenient compared to *Base*.^{18,19}

5.1.2. Grade Distributions

Figure 3 reports the five grade distributions that are selected most often by the supervisors in each treatment (see Table A2 for an overview of all grade distributions). For example, the distribution (1,2,3) refers to a situation where one worker is graded with a “1”, another worker receives grade “2”, and the third worker gets grade “3”. Interestingly, Figure 3 shows that under each rating policy another grade distribution is chosen most often.

In *Base*, the distribution (1,1,1) is most frequently chosen: In approximately 27 % of the observations, all three workers are simultaneously rated with the best grade by their supervisor. In *Rec*, the distribution (1,2,3) stands out with 28 %, and in *Bud*, the distribution (2,3,4) is prominent with 42 %. Comparing the distributions that are most often selected, it is visible that the recommendation to differentiate and the financial budget lead to less lenient and more differentiated ratings compared to *Base*. In *Rec*, in only 7 % of the observations, all three workers simultaneously obtain the best grade by their supervisor. This amounts to approximately 20 percentage points less compared to *Base*. Please note that the (compressed) distribution (3,3,3) is the second most common distribution in *Bud*. Hence, it cannot be clearly argued that the ratings in *Bud* are more differentiated compared to *Base*.

5.1.3. Relative Performance Standards

Furthermore, it is questionable whether the allocation of performance appraisals is in line with the relative performance of workers, i.e., whether the high-performing workers receive top grades, and the low-performing workers obtain bad grades.²¹ We observe from

¹⁸ Figures A1, A2 and A3 in the appendix show grade distributions for each period in *Base*, *Rec* and *Bud* to enhance understanding of how grades are allocated in each period.

¹⁹ The Kolmogorov-Smirnov test shows that the distributions of grades differ significantly between *Rec* and *Base* ($D = 0.21$, $p\text{-value} = 0.00$) and between *Bud* and *Base* ($D = 0.53$, $p\text{-value} = 0.00$).

²⁰ In Figure 3, group-level observations are included.

²¹ There are only three observations where all three workers generate exactly the same output, and in approximately 15 % of all observations two workers have the same output within their group. Hence, in the major part of all observations all three workers differ in their performance. Also the standard deviation of output signals that workers vary with their performance.

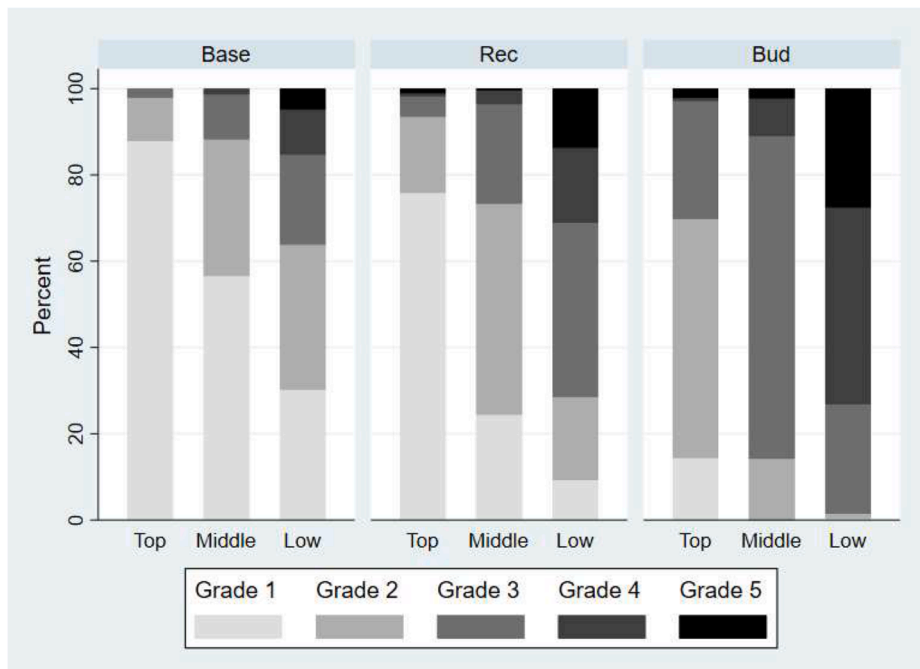


Figure 4. Grade Distribution over Top, Middle and Low Performer Periods 1-16²³.

Figure 4 that still about 60 % of the low performers²² in *Base* are graded with one of the good grades "1" or "2". Compared to *Base*, only about 25 % of the low performers receive a good grade in *Rec*. Instead, they are more often assigned the grade "3", "4" or "5". Also in *Bud*, more low performers receive bad grades compared to *Base*. Consequently, both rating policies help to achieve less lenient performance appraisals relative to *Base*.

Figure 4 already pictures that the relative performance is given greater consideration in *Rec* and *Bud* compared to *Base*. This finding is further strengthened by the regression analyses depicted in Table 3. Across all regression analyses – columns (1) to (5) – the dependent variable is the individual bonus a worker receives. The independent variables, *Rank1* and *Rank2*, are dummy variables and represent the workers' rank within their group. *Rank1* equals one if the worker is a top performer, and *Rank2* equals one if the worker is a middle performer. The reference category, *Rank3*, represents low performer.

In all treatments, the individual output positively impacts the individual bonus. Columns (1) to (3) indicate that top performers (*Rank1*) receive a higher individual bonus than low performers (*Rank3*). Similarly, middle performers also receive a higher individual bonus compared to low performers. In each treatment, supervisors seem to take relative performance into account. Since the coefficients are larger in *Rec* (column 2) and *Bud* (column 3) compared to *Base* (column 1), we assume that the consideration of relative performance is more pronounced when the rating policies are present.

Column (4) shows that the interaction of *Rank1* and the recommendation to differentiate is weakly significant. Thus, rank one is slightly more important than rank three under *Rec* compared to *Base*: When supervisors receive the recommendation to differentiate, the top performers are rewarded with higher bonus payments, and therefore better grades, compared to receiving no recommendation. This effect is even stronger for *Bud*, where the interaction term is strongly significant. Additionally, rank two also has a significant positive impact on the individual bonus compared to rank three under the budget constraint. Overall, the rating policies lead supervisors to consider relative performance standards more strongly.^{24,25}

5.1.4. Degree of Differentiation

Moreover, we measure the degree of differentiation by calculating the span of grades, defined as the difference between the worst

²² The top, middle and low performers are identified by ranking the workers' output in their group in each period. Hence, the worker with the highest output in a group is the top performer, and the worker with the lowest output is the low performer. In case of a draw, a tie-breaker rule is applied.

²³ In Figure 4, individual-level observations of workers are included.

²⁴ When assessing the interaction effects for each part separately, it can be seen that this result applies equally to part one and two of the experiment.

²⁵ Table A3 shows that an individual's share of group output (i.e., individual output relative to group output) is positively associated with an individual's share of group bonus payments (i.e., individual bonus relative to bonus payments received by all group members) across all treatments.

and the best grade assigned by the supervisor within a group in one period. A high span of grades indicates a high degree of differentiation, whereas a low span of grades reflects a low degree of differentiation. If the span of grades equals zero, there is no differentiation because all three workers receive the same grade. The average span of grades of *Rec* (1.86) and *Bud* (1.97) is larger than the average span of grades of *Base* (1.23). The MWU test states a significant difference between *Rec* and *Base* ($z = -2.81$, $p\text{-value} = 0.00$) and also between *Bud* and *Base* ($z = -2.70$, $p\text{-value} = 0.01$).^{26, 27, 28}

The regression analyses in Table 4 indicate that the recommendation to differentiate leads to a significantly higher span of grades compared to *Base* in both parts of the experiment. Again, *Hypothesis 1a* can be confirmed. Also the budget leads to a significantly higher span of grades which can be seen in Table A4 in the appendix. This contradicts *Hypothesis 1c*, where we expected less differentiated appraisals under a budget compared to *Base*. In the regression analyses, we include control variables such as the standard deviation of the output and the group output. We control for the standard deviation of workers' output because we assume that a high standard deviation, i.e., a heterogeneous performance within a group, may lead to more differentiated performance appraisals, which the regression results support. Group output is also controlled for since a high group output likely indicates strong performance across all workers. In such case, supervisors may tend to allocate similarly good grades to the three workers, resulting in a lower span of grades. This is also in line with the regression results.²⁹

The regression analyses in Table 5 demonstrate how the degree of differentiation changes over the course of the experiment while controlling for the standard deviation of the workers' output and the group output. The dependent variable is the span of grades as previously defined. The time is modelled by the four dummy variables *Periods 1-4*, *Periods 5-8*, *Periods 9-12* and *Periods 13-16*, whereas *Periods 1-4* constitutes the reference category. The first column of Table 5 shows that the increase in the span of grades in *Periods 5-8*, *Periods 9-12* and *Periods 13-16* compared to the first four periods is not significant in *Base*. Column (2) presents a significant decrease in the span of grades in *Periods 9-12* and a weakly significant decrease in *Periods 13-16* compared to the first four periods in *Rec*. In *Bud*, the regression analysis does not show a clear time trend.

5.1.5. Insights into Supervisors' Behavioral Drivers

The post-experimental questionnaire offers additional insights into supervisors' motives (see Table A5 in the appendix). Supervisors rate each item separately for the first and second part of the experiment on a seven-point scale, ranging from "does not apply at all" (1) to "fully applies" (7). The following analysis focuses on the *Bud* treatment focusing on the most informative patterns emerging in the questionnaire.

Results show a significantly higher span of grades in *Bud* than in *Base*. This is further supported by the item - "I made use of the whole range of grades" - which is on average higher in *Bud* than in *Base* (part one: $z = -2.76$, $p\text{-value} = 0.00$; part two: $z = -3.47$, $p\text{-value} = 0.00$).

As outlined in the hypothesis section, supervisors in *Bud* face a trade-off between assigning equal ratings and differentiated rating; assigning a good grade necessitates giving a low grade as well. As a result, the budget restriction may drive supervisors to make stricter decisions. They may assess workers' performance harsher, assigning bad grades to low performers more quickly, thereby adhering to the budget restriction in order to reward top-performing workers. This explanatory approach is supported by the post-experimental questionnaire. Especially, in the second part of *Bud*, supervisors strongly confirm the statement: "I gave negative evaluations to punish the workers". Moreover, supervisors state that relative comparisons of workers' performance played a more important role in *Bud* than in *Base*. This suggests that punitive evaluations may be a consequence of the budget restriction.

There are some supervisors that have the tendency to assign equal ratings (grade "3" to all workers). The item "It would have been fairer, or it was the fairest, to give all workers the same payment regardless of their performance" does not differ significantly between *Bud* and *Base*, indicating that egalitarian preferences remain unaffected by the budget constraint. However, the form of equal ratings differs: In *Base* supervisors rather assign grade "1" to all three workers while in *Bud* grade "3" is more common.

Moreover, supervisors in both *Bud* and *Base* feel similarly capable of making evaluation decision and perceive their ratings as equally fair across treatments. Though, supervisors feel significantly more uncomfortable to evaluate workers in *Bud*, likely reflecting the added pressure of managing trade-offs under the budget constraint.

Interestingly, there are almost no significant differences between *Rec* and *Base* in the questionnaire. This suggests that supervisors in *Rec* felt similarly to those in *Base* regarding their performance appraisals.

²⁶ The average span of grades across all 16 periods for each supervisor is calculated, resulting in an average value for each supervisor. These averages are compared between *Base* and *Rec/Bud*.

²⁷ Also the two-sided t-test indicates a significant difference between *Rec* and *Base*, when comparing averages ($p\text{-value} = 0.00$, observed treatment effect = 0.64). Given the limited number of independent observations per treatment, statistical power may be constrained. To provide a more precise idea of our results' meaning, we compute the minimum detectable effect (MDE) using Stata's *power twomeans* command ($\text{power} = 0.8$, $\alpha = 0.05$). Specifically, we first calculate the average span of grades across all 16 periods for each supervisor, and then use the standard deviations of these averages to determine the MDE. The resulting MDE of 0.6 is smaller than the observed treatment effect of 0.64. This suggests that the observed effect is robust given the study's design.

²⁸ The two-sided t-test reveals a significant difference between *Bud* and *Base* ($p\text{-value} = 0.00$, observed treatment effect = 0.74). The respective minimum detectable effect of 0.6 is smaller than the observed treatment effect of 0.74.

²⁹ In columns (2) and (4) of Table 4, we find no evidence that selfish, envious or equity-oriented supervisors evaluate differently compared to altruistic supervisors.

Table 3

Impact of relative performance on individual bonus (periods 1–16).

Dependent Variable:	Individual Bonus				
	(1) <i>Base</i>	(2) <i>Rec</i>	(3) <i>Bud</i>	(4) <i>Base vs. Rec</i>	(5) <i>Base vs. Bud</i>
Output	0.267*** (0.032)	0.247*** (0.029)	0.125*** (0.010)	0.269*** (0.029)	0.239*** (0.029)
Output $\times$ Rec				−0.021 (0.039)	
Rank2	0.762*** (0.208)	1.167*** (0.226)	1.602*** (0.189)	0.766*** (0.203)	0.841*** (0.201)
Rank1	1.037*** (0.330)	1.996*** (0.365)	2.957*** (0.282)	1.052*** (0.324)	1.180*** (0.316)
Rank2 $\times$ Rec				0.390 (0.300)	
Rank1 $\times$ Rec				0.926* (0.483)	
Rec				−1.415* (0.823)	
Output $\times$ Bud					−0.108*** (0.029)
Rank2 $\times$ Bud					0.652** (0.274)
Rank1 $\times$ Bud					1.610*** (0.429)
Bud					−2.051*** (0.726)
Individual Ability	−0.100** (0.051)	−0.161*** (0.048)	−0.069*** (0.022)	−0.131*** (0.037)	−0.077*** (0.027)
Constant	4.426*** (0.914)	3.624*** (0.732)	1.983*** (0.522)	4.714*** (0.795)	4.306*** (0.705)
Observations	768	768	768	1536	1536
Number of Subjects	48	48	48	96	96
Wald Chi2	-	-	-	1296.00	3783.00

Random Effects Regression (period dummies included). Reference category: Rank3 and *Base*.

Columns (1), (2) and (3): Observations are based on 48 workers and 16 periods.

Columns (4) and (5): Observations are based on 48 workers, 16 periods and 2 treatments.

Robust standard errors in parentheses (clustered on group_id). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.2. Impact of Rating Policies on Performance

The second research aim of this paper is to investigate how the rating policies or rather the supervisors' rating behavior impact workers' performance. Accordingly, we first examine whether workers perform differently between treatments. Moreover, we provide insights into workers' behavioral drivers by analyzing the post-experimental questionnaire.

5.2.1. Workers' Performance

To clarify the measures used, the *group output* is defined as the sum of individual outputs per group. This measure can be averaged in two ways: Across all groups for one period (*average group output per period*) or across all periods for one group (*average group output per group*).

Figure 5 illustrates the average group output per period over the course of the experiment.³⁰ In each period of the first part, the average group output per period in *Rec* and *Bud* exceeds the values in *Base*. Although no significant difference between *Rec* and *Base* emerges from the MWU test for the first part ($z = -1.36$, $p\text{-value} = 0.18$),³¹ later regressions show a significant difference when

³⁰ To test whether workers perform equally across treatments in the preliminary period, we conduct the MWU test to compare individual outputs. There is no significant difference between *Rec* and *Base* (MWU test; $z = -0.23$, $p\text{-value} = 0.82$), nor between *Bud* and *Base* (MWU test; $z = -1.13$, $p\text{-value} = 0.26$). Accordingly, the randomization regarding subjects' ability worked well.

³¹ The group outputs are averaged across periods from the first part, resulting in an average value for each group. These group averages are compared between *Base* and *Rec/Bud*.

Table 4
Impact of recommendation to differentiate on span of grades.

Dependent Variable:	Span of Grades			
	(1) Part 1	(2) Part 1	(3) Part 2	(4) Part 2
<i>Rec</i>	0.697*** (0.206)	0.544** (0.246)	0.507** (0.220)	0.526* (0.294)
SD of Output	0.170*** (0.024)	0.174*** (0.025)	0.187*** (0.018)	0.186*** (0.019)
Group Output	−0.035*** (0.008)	−0.035*** (0.009)	−0.019** (0.008)	−0.021** (0.010)
Male		−0.079 (0.234)		−0.340 (0.219)
Age		−0.017 (0.051)		−0.071 (0.059)
Economics		−0.139 (0.203)		−0.206 (0.312)
Selfish		0.478 (0.330)		0.243 (0.268)
Envy		−0.148 (0.397)		0.0708 (0.398)
Equity		−0.109 (0.275)		0.0858 (0.311)
Constant	2.484*** (0.470)	3.647* (1.959)	1.765*** (0.636)	3.549 (2.535)
Observations	256	248	256	248
Number of Subjects	32	31	32	31
Wald Chi2	155.48	228.09	212.81	267.00

Random Effects Regression (period dummies included). Reference category: *Base* and *Altruism*.

Observations are based on 16 supervisors and 16 periods.

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 5
Degree of differentiation over time.

Dependent Variable:	Span of Grades		
	(1) <i>Base</i>	(2) <i>Rec</i>	(3) <i>Bud</i>
Periods 5–8	0.075 (0.175)	−0.099 (0.149)	0.064 (0.226)
Periods 9–12	0.026 (0.171)	−0.289** (0.146)	0.129 (0.158)
Periods 13–16	0.043 (0.233)	−0.347* (0.181)	−0.054 (0.177)
SD of Output	0.195*** (0.032)	0.186*** (0.021)	0.152*** (0.018)
Group Output	−0.038*** (0.008)	−0.009 (0.007)	−0.007*** (0.003)
Constant	2.735*** (0.423)	1.675*** (0.437)	1.492*** (0.280)
Observations	256	256	256
Number of Subjects	16	16	16
Wald Chi ²	69.11	123.44	376.42

Random Effects Regression. Reference category: Periods 1–4.

Observations are based on 16 supervisors and 16 periods.

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

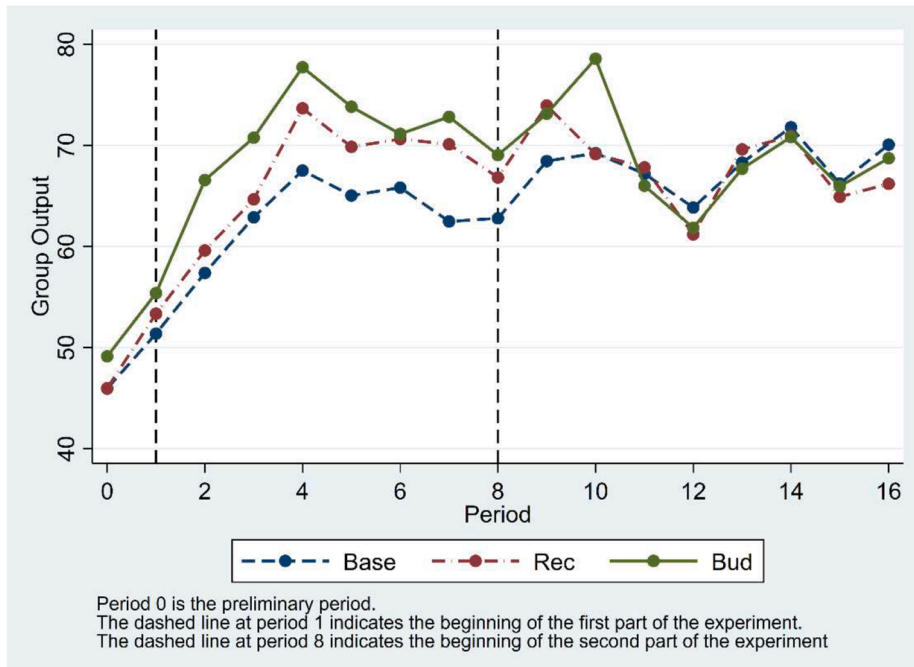


Figure 5. Group Output from Periods 1-16³⁵.

controlling for ability.³² Considering *Bud* and *Base*, the performance difference is significant in the first part (MWU test, $z = -1.88$, $p\text{-value} = 0.06$).³³ In the second part, the performance of *Rec* and *Bud* both adjust to the performance of *Base*.³⁴ The MWU test reveals that there is no significant performance difference between *Rec* and *Base* ($z = 0.08$, $p\text{-value} = 0.96$) or *Bud* and *Base* in the second part ($z = -0.66$, $p\text{-value} = 0.52$).

We also ran several regression analyses to investigate whether the workers' performance differs significantly in the first part (see Table 6 and Table 7). According to the OLS regression (first column of Table 6), the average group output per group is approximately 4.2 points higher in *Rec* compared to *Base* in the first part of the experiment when we control for the group ability. Also the random effects regressions (columns 2 and 3) indicate that recommending differentiated performance appraisals lead to a significantly higher performance compared to giving no restrictions. When the logarithm of the output is used as the dependent variable, it can be observed that *Rec* leads to an approximately 6.5 percent higher average performance compared to *Base* (column 4). When adding further controls, the performance difference is still significant (see Table A7). When subjects are male, the average output increases by 5 percent, but this effect is only weakly significant. There is no significant gender effect if the dependent variable is the average output (not logarithmized).³⁶

The first column of Table 7 shows that the average group output per group is approximately 5.6 points higher in *Bud* compared to *Base* in the first part of the experiment while controlling for group ability. This corresponds to an approximately 8.1 percent higher performance in *Bud* compared with *Base*. The random effects regressions also support this result. The performance is still significantly higher in *Bud* compared to *Base* when adding further controls (see Table A8 columns 2 and 5).

All in all, the results show that the performance in the first part of the experiment is significantly higher in *Rec* compared to *Base* when controlling for workers' ability. Also in *Bud*, workers' performance is significantly higher compared to *Base* in the first part. In the

³² The two-sided t-test indicates no significant performance difference between *Rec* and *Base*, when comparing group averages in the first part ($p\text{-value} = 0.26$, observed treatment effect = 4.18). Using the procedure in footnote 27, the minimum detectable effect for the first part is 10.45 which is higher than the observed treatment effect of 4.18. This suggests that the study is powered to detect larger effects, meaning that a larger sample size may lead to significant performance differences.

³³ The two-sided t-test indicates a weakly significant performance difference between *Bud* and *Base*, when comparing group averages in the first part ($p\text{-value} = 0.06$, observed treatment effect = 7.75). Using the procedure in footnote 27, the minimum detectable effect for the first part is 11.66 which is higher than the observed treatment effect of 7.75. This suggests that our design is powered to detect larger effects, so moderate effects like the one observed are expected to be detected less consistently. The results provide evidence in favor of *Bud* and motivate a modest increase in sample size in future research.

³⁴ The curves follow a similar pattern across treatment. This pattern may reflect learning or fatigue effects, as no external variables were introduced that could otherwise explain such trends.

³⁵ In Figure 5, group-level observations are included.

³⁶ In further analyses, we find evidence that male workers react more strongly to *Rec* than female workers by showing a larger performance increase. Since the experiment is not designed to analyze gender differences, we do not elaborate on this result further.

Table 6

Regression analyses presenting the impact of recommendation to differentiate on output part one.

Dependent Variable:	Output			Log Output		
	Base vs. Rec (periods 1–8)			Base vs. Rec (periods 1–8)		
	OLS	RE (Groups)	RE (Individuals)	OLS	RE (Groups)	RE (Individuals)
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Rec</i>	4.156** (1.812)	4.156** (1.785)	1.386** (0.588)	0.065** (0.028)	0.065** (0.028)	0.054** (0.027)
Group Ability	0.617*** (0.068)	0.617*** (0.067)		0.010*** (0.001)	0.010*** (0.001)	
Individual Ability			0.604*** (0.065)			0.030*** (0.003)
Constant	33.540*** (3.131)	21.910*** (3.112)	7.514*** (1.002)	3.660*** (0.050)	3.463*** (0.050)	2.339*** (0.048)
Observations	32	256	768	32	256	758
Number of Groups (Subjects)	32	32	96	32	32	96
R ² / Wald Chi ²	0.77	585.09	668.24	0.77	750.48	608.45

The random effect regressions include period dummies. Reference category: *Base*. Columns (1) and (4): Observations are based on 16 groups and 2 treatments.

Columns (2) and (5): Observations are based on 16 groups, 2 treatments and 8 periods. Columns (3) and (6): Observations are based on 48 workers, 2 treatments and 8 periods (in (6), 10 workers with an output of zero are excluded since the logarithm is undefined for zero). Robust standard errors in parentheses (in (3) and (6) clustered on group_id).

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

second part, there are no significant performance differences between *Rec* or *Bud* and *Base* according to the non-parametric results and the regression analyses.

There are several possible reasons why the performance difference vanishes in the second part. The supervisors may have changed their rating behavior in the second part which could cause a change in workers' performance. As described in chapter 5.1.4, the supervisors differentiate less in the second part in *Rec* when controlling for the standard deviation of output. This could be one reason for the decreasing performance in *Rec*.

Another cause of the declining performance could be rooted in the frequency of pushing the timeout button as a consequence of workers not being motivated. As soon as one worker pushes the timeout button, the worker's screen locks for 20 seconds, and the worker is not able to count sevens anymore which leads to a stagnating performance. It can be found in Figure A4 in the appendix that the average group timeouts per period³⁷ in *Rec* and *Bud* are above those of *Base* in the second part. The difference is not significant for *Rec* and *Base* (MWU test, $z = -0.39$, p -value = 0.71), but it differs significantly between *Bud* and *Base* ($z = -2.88$, p -value = 0.00).³⁸ The higher timeouts in the second part can be partially explained by a few workers who exhibit an extreme behavior: These workers almost completely stop their performance in the second part and continuously click the timeout button instead.³⁹

Middle/low performers may take more timeouts because they have to fear bad grades and a low payoff which may be demotivating. In Table A9 in the appendix, it is visible that the low performers significantly more often push the timeout button compared to the top performers in *Rec* and *Bud*. However, this result rather shows a correlation than a causal relationship.⁴⁰ Table A10 in the appendix presents that middle and low performers in *Rec* increase the timeouts when receiving a bad grade in the second part. Though, this behavior cannot be observed for the first part. In *Bud*, middle and low performers receiving bad grades choose the timeouts more often in both parts of the experiment (see Table A11 in the appendix). Thus, pressing the timeout button seems to be a reaction to bad grades.

5.2.2. Differentiation and Performance

The results show that both rating policies help to significantly increase performance in the first part of the experiment. Next, we analyze the underlying mechanisms that may cause a higher performance when the rating policies are applied. One possible explanation for the performance differences could be that differentiated ratings in *Rec* or *Bud* lead to a higher performance. First, we examine this assumption for *Rec*.

Whether a higher degree of differentiation leads to a higher subsequent performance, can be observed in Table 8. The table shows random effects regressions where the group output of the subsequent period is the dependent variable. Again, the span of grades functions as a measure for differentiation and is included as dummy variables. The reference category is the span of grades of zero, i.e.,

³⁷ The group timeouts are defined as the sum of individual timeouts per group. This measure can be averaged in two ways: Across all groups for one period (average group timeouts per period) or across all periods for one group (average group timeouts per group).

³⁸ The average group timeouts per group are compared between *Base* and *Rec/Bud*. In the first part, there is no significant difference between *Base* and *Rec/Bud* (MWU test, $z = -0.35$, p -value = 0.74 and $z = -1.17$, p -value = 0.25).

³⁹ Within 2.5 minutes, the timeout button can be pressed for seven times at most which leads to a payoff of 1.75 €.

⁴⁰ It is not clear whether low-performing workers tend to press the timeout button, or whether pushing the timeout button leads to being a low performer.

Table 7

Regression analyses presenting the impact of budget on output part one.

Dependent Variable:	Output			Log Output		
	Base vs. Bud (periods 1–8)			Base vs. Bud (periods 1–8)		
	OLS	RE (Groups)	RE (Individuals)	OLS	RE (Groups)	RE (Individuals)
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Bud</i>	5.627** (2.364)	5.627** (2.328)	1.856** (0.766)	0.0814** (0.035)	0.081** (0.035)	0.077** (0.032)
Group Ability	0.665*** (0.079)	0.665*** (0.078)		0.010*** (0.001)	0.010*** (0.001)	
Individual Ability			0.683*** (0.076)			0.032*** (0.003)
Constant	31.360*** (3.726)	18.960*** (3.822)	6.038*** (1.252)	3.446*** (0.054)	3.446*** (0.054)	2.297*** (0.051)
Observations	32	256	768	256	256	752
Number of Groups (Subjects)	32	32	96	32	32	96
R ² / Wald Chi ²	0.71	382.70	403.10	0.72	460.11	481.67

The random effect regressions include period dummies. Reference category: *Base*. Columns (1) and (4): Observations are based on 16 groups and 2 treatments.

Columns (2) and (5): Observations are based on 16 groups, 2 treatments and 8 periods. Columns (3) and (6): Observations are based on 48 workers, 2 treatments and 8 periods (in (6), 16 workers with an output of zero are excluded since the logarithm is undefined for zero). Robust standard errors in parentheses (in (3) and (6) clustered on group_id).

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

all three workers are graded with the same grade. The group output, the standard deviation of the output and the group ability are included as control variables.

The standard deviation of the output within one group is added as a control variable because workers are aware of their co-workers' performance which may influence their effort. We argue that workers may perceive the recommendation to differentiate as a competitive situation: They must assume that supervisors follow the recommendation, leading workers to compete for the good grades.⁴¹ According to tournament theory, workers are expected to exert the highest effort in tournaments with workers that are homogenous in abilities (Lazear & Rosen, 1981). In our experiment, performance in the previous period may serve as an indicator for workers' abilities. Therefore, a low standard deviation of the group output, indicating homogeneous abilities, is expected to lead to higher subsequent performance. Additionally, it is controlled for the group ability because groups with a higher ability are likely to generate a higher output. Except for the standard deviation of the output, the controls have a significant impact on the subsequent group output.

According to column (1) of Table 8, the span of grades of three leads to a significantly higher group output in the subsequent period compared to no differentiation (span of grades of zero). Columns (2) and (3) demonstrate that this result applies only to the first part of the experiment because the coefficient, *Span of Grades*=3₁, is only significant in the first part. In the first part, also a span of grades of four leads to a significantly higher group output in the subsequent period compared to a span of grades of zero. To conclude, a high degree of differentiation leads to a higher subsequent group output compared to no differentiation which is in line with *Hypothesis 2a*. This result applies to the first part of *Rec*. In Table A12, we examine whether top, middle and low performers react differently towards differentiated performance appraisals. It can be observed that especially low performers significantly increase their performance when there is a high degree of differentiation compared to no differentiation in the first part of the experiment (see column 4).⁴²

In contrast to *Rec*, a higher degree of differentiation leads to a smaller subsequent group output in *Bud*. Column (4) of Table 8 shows that a span of grades of four leads to a significantly smaller group output in the following period compared to no differentiation (*Span of Grades*=0₁). This effect is driven by the first part of the experiment because there are no significant effects in the second part (see column 6). It can be observed from column (5) that a span of grades of three and four lead to a significantly smaller subsequent group output compared to no differentiation. In *Bud*, a span of grades of zero mostly contains the grade distribution (3,3,3) (see Table A2). Hence, assigning grade three to all workers leads to a significantly higher performance compared to differentiated performance appraisals. Therefore, *Hypothesis 2b* cannot be confirmed. Again, we analyze whether top, middle and low performers react differently towards differentiated appraisals. Column (4) of Table A13 presents that spans of grades amounting to two, three and four have a significant, negative impact on the subsequent individual output compared to no differentiation for low performers. In contrast to low performers, top performers feature a different behavior. A span of grades of three leads to a significantly positive individual output compared to no differentiation.

⁴¹ The budget restriction may also create a competitive environment, as workers know that supervisors are facing a trade-off between assigning good and bad grades. Therefore, workers may compete for the good grades.

⁴² This result stays robust if we add further controls such as the subjects' gender, age, personality traits, positive/negative reciprocity, whether subject studies economics and whether subject perceives fun.

Table 8Impact of span of grades on subsequent group output in *Rec* and *Bud*.

Dependent Variable:	Group Output _{t+1}					
	<i>Rec</i>			<i>Bud</i>		
	(1) Periods 1–16	(2) Periods 1–8	(3) Periods 9–16	(4) Periods 1–16	(5) Periods 1–8	(6) Periods 9–16
Span of Grades=1 _t	2.408 (2.764)	5.429 (3.920)	−1.008 (4.088)	1.091 (4.806)	−0.422 (8.537)	3.278 (6.464)
Span of Grades=2 _t	0.702 (2.559)	3.280 (4.169)	−2.088 (1.931)	−2.314 (2.880)	−5.237 (3.616)	1.265 (3.888)
Span of Grades=3 _t	5.688* (3.062)	7.542* (4.056)	3.591 (4.209)	−4.240 (3.795)	−8.204** (4.039)	−0.604 (6.001)
Span of Grades=4 _t	5.468 (3.525)	10.720** (4.810)	0.682 (5.090)	−9.917** (4.625)	−11.200** (4.870)	−6.195 (5.564)
Group Output _t	0.647*** (0.087)	0.591*** (0.119)	0.685*** (0.109)	0.528*** (0.066)	0.451*** (0.131)	0.533*** (0.072)
SD of Output _t	−0.428 (0.304)	−0.510 (0.377)	−0.293 (0.365)	0.844** (0.377)	1.311** (0.555)	0.500 (0.530)
Group Ability	0.214*** (0.079)	0.244** (0.098)	0.177 (0.109)	0.288** (0.128)	0.426*** (0.153)	0.205 (0.175)
Constant	14.620*** (4.797)	14.000*** (4.770)	13.000** (5.286)	22.110*** (6.714)	20.350*** (6.426)	26.250*** (9.334)
Observations	240	128	112	240	128	112
Number of Groups	16	16	16	16	16	16
Wald Chi ²	-	6979.25	-	-	367.67	-

Random Effects Regression (period dummies included). Reference category: Span of Grades = 0_t Observations are derived from group and period combinations.

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.2.3. Insights into Workers' Behavioral Drivers

Table A6 in the appendix summarizes workers' responses to the questionnaire. In both *Rec* and *Bud*, workers' motivation declines from the first to the second part. For example, agreements with the items "I put a lot of effort into the task" or "I really wanted to get the best evaluation in my group" are significantly lower in the second part. Similarly, the motivational effect of positive or negative evaluations diminishes in the second part. These findings support the notion that the performance decrease in the second part is partly attributable to reduced motivation.

In both *Bud* and *Rec*, workers evaluate supervisors less favorably than in *Base* whereby workers seem to be more dissatisfied in *Bud*.⁴³ For example, workers rate the statement – "I have the impression that the intensity of my efforts played an important role in my evaluations" – higher in *Base* compared to *Bud* and *Rec*.⁴⁴ Similarly, the item "I consider my supervisor capable of making appropriate evaluations" receives lower agreement in *Bud* and *Rec* compared to *Base*, particularly in the second part. This suggests that workers are more satisfied with evaluations and supervisors when no specific rating policy is implemented.

Moreover, workers perceive the evaluations in *Base* as fairer compared to *Bud*.⁴⁵ and trust in supervisors is significantly lower.⁴⁶ This is evident from the item "I would feel comfortable letting my supervisor control my payoff in the future as fell". Overall, perceived fairness and trust are particularly compromised in *Bud*.

5.3. Costs of Rating Policies

As previously presented, both rating policies lead to a significantly higher performance in the first part of the experiment. From an organizational perspective, the relationship between costs and benefits of the respective rating policy is important to decide which rating policy may be the most efficient one. In this case, "benefit" refers to a worker's output, and "costs" refers to the bonus payment. Consequently, we create the measure: Costs per Output Unit = $\frac{\text{Bonus Payment}}{\text{Output}}$.⁴⁷ According to Figure A5 in the appendix, the costs per output unit are smallest in *Bud* across all periods. The difference between *Bud* and *Base* is significant for the first and the second part of the experiment (MWU test; Part 1: $z = 4.75$, $p\text{-value} = 0.00$; Part 2: $z = 4.33$, $p\text{-value} = 0.00$).⁴⁸ Also the difference between *Bud* and *Rec* is significant for each part (MWU test; part 1: $z = -3.81$, $p\text{-value} = 0.00$; part 2: $z = -3.05$, $p\text{-value} = 0.00$). Consequently, *Bud* has a

⁴³ In *Bud*, there are more statements concerning the supervisor that exhibit significant differences from *Base*.

⁴⁴ The difference to *Base* is statistically significant.

⁴⁵ This aspect was measured via the item: „My supervisor's behavior towards the workers was fair.“

⁴⁶ Workers may perceive the evaluations in *Bud* as less fair than in *Base* because receiving a good grade in *Bud* entails negative consequences for other group members.

⁴⁷ The ratio states how much money is spent for one output unit of a worker. Thus, the smaller the ratio, the better is the relationship between output and bonus payment.

⁴⁸ The average costs per output unit per group of *Bud* are compared to those of *Base* for each part of the experiment separately.

better cost-benefit ratio compared to *Base* and compared to *Rec*. Also the difference between *Rec* and *Base* is significant in the first and second part (MWU test; part 1: $z = 2.60$, $p\text{-value} = 0.01$; part 2: $z = 2.00$, $p\text{-value} = 0.05$). To sum up, the cost-benefit ratio of *Bud* is the best, but also in *Rec* the cost-benefit ratio is significantly better compared to *Base*.

6. Discussion

In this section, we interpret the experimental results. First, we discuss supervisors' rating behavior. Then, we elaborate on potential drivers of performance differences, analyzing the relationship between differentiated ratings and worker performance in more detail. Our design allows to explicitly test whether differentiated ratings operate as a mechanism for increased performance. However, other mechanisms cannot be directly examined. Hence, we propose plausible assumptions regarding these mechanisms, leaving their verification for future research.

The analysis of supervisors' rating behavior reveals that the recommendation to differentiate results in less lenient and more differentiated performance appraisals compared to *Base*. Contrary to our expectations, the span of grades in *Bud* is significantly higher compared to *Base*. However, it is evident that supervisors tend to assign grade "3" to all three workers simultaneously in *Bud*, aligning with our expectation. Those supervisors seem to avoid giving bad grades and rather assign similarly "good" grades to all three workers.

Furthermore, the experiment demonstrates that both rating policies lead to an increasing performance (at least in the first part). One possible explanation for the higher performance is the higher degree of differentiation. Both rating policies lead to a higher span of grades, suggesting that the differentiation itself may contribute to the increase in performance.

Indeed, our results show that a higher degree of differentiation leads to a higher performance in the first part of *Rec*. This effect appears to be primarily driven by low-performing workers. In *Rec*, a span of grades of zero mostly includes the grade distribution (1,1,1) (see Table A2). When all workers are graded with the best grade simultaneously, their actual performance may not be properly considered in the grading, i.e., also the middle and low performer receive the best grade. In contrast to a scenario with low differentiation, a high span of grades leads to mediocre or bad grades for middle and low performers, while top performers continue to receive good grades. As depicted by Figure 4 in section 5.1.3, bad grades are almost exclusively assigned to low performers, making them the primary recipients of bad evaluations in differentiated settings. Since a high degree of differentiation especially affects the grades of middle and low performers, it is likely that the performance increase is driven by them.⁴⁹ Consequently, when there is a high degree of differentiation, low performers are the ones who increase their performance which may be a reaction to bad grades.⁵⁰ Low performers reacting positively towards bad grades is in line with our assumptions: Differentiated ratings lead to a higher marginal return to effort compared to compressed ratings, making it more rewarding for workers to increase their effort.

In *Bud*, the opposite effect emerges: A high degree of differentiation leads to a lower subsequent performance. This effect is driven by low performers who decrease their performance when there is a high span of grades. In *Bud*, supervisors assign bad grades almost exclusively to low-performing workers (see Figure 4). Consequently, low performers are likely to receive bad grades under a high span of grades, and their reaction to a high span of grades, or rather a bad grade, appears to be negative compared to receiving a grade of "3."

Several factors may explain this behavior. While workers are unaware of their co-workers' bonus payments, they can infer inequalities due to the budget restriction. Low performers may recognize that equal bonus payments are only achievable if all workers receive a grade of "3" (5 €). Thus, if they receive a grade of "3", they may assume that others do as well. For low performers, such an assumption may elicit positive reciprocity, as they feel fairly treated because they prefer equal ratings.

Positive reciprocity may also arise from gratitude for not being punished with a bad grade, thereby motivating workers to enhance their performance.⁵¹ Conversely, negative reciprocity may explain their negative reaction towards a high degree of differentiation. Dissatisfaction with a bad grade, may lead to a reduced performance. Furthermore, low performers may be affected by feelings of perceived fairness.⁵² There is empirical evidence that individuals tend to overestimate their performance (Krueger & Mueller, 2002; Meyer, 1975). Consequently, low performers may overestimate their actual performance and may perceive grade "3" as fairer compared to a bad grade. In each period, workers observe their performance relative to their co-workers which can lead to changes in their perception of abilities compared to others throughout the experiment.

In contrast to low performers, top performers feature a different behavior: A higher degree of differentiation (span of grades of three) leads to a significantly higher individual output compared to no differentiation. Under high differentiation, top performers are rewarded with good grades, whereas in the absence of differentiation, they receive grade "3." Thus, being rewarded with a good grade seems to be motivating for top performers.⁵³

⁴⁹ Top performers are not directly affected by a low or high degree of differentiation because they receive good grades anyway.

⁵⁰ When there is a span of grades of four, the low performers always receive grade "5", and if the span of grades equals three, the low performers always receive grade "4" or "5" in the first part in *Rec*.

⁵¹ There is experimental and field evidence that higher payoffs may be rewarded with a high performance due to positive reciprocity. See for example: Charness (2004), Fehr et al., (1997) or Kube et al., (2012).

⁵² The perceived fairness of payments may have a significant impact on performance (Cohen-Charash & Spector, 2001).

⁵³ Also these results stay robust when adding further controls: gender, age, personality traits, positive/negative reciprocity, whether subject studies economics and whether subject perceives fun.

It is questionable why differentiated ratings have a different effect on performance in *Rec* and *Bud*. Since the recommendation to differentiate must not necessarily be followed by the supervisor, low performers know that every group member (including themselves) can potentially reach a good grade. In *Bud*, low performers know that they need to outperform the top performer in order to reach a good grade. This substantial performance increase that is needed to obtain a good grade may destroy performance incentives for low performers. Another explanatory approach is that the grade distribution (3,3,3) works particularly well in *Bud*. As previously described, low performers may appreciate receiving grade “3” because they are positively reciprocal or perceive it as more fair.

The performance differences wear off over the course of the experiment. The performance decrease may be driven by a few individuals that show an extreme behavior by stopping to work and pressing the timeout button. Another reason could be that supervisors stop differentiating in the second part of the experiment. In *Rec*, supervisors differentiate less in the second part compared to the first part when controlling for the workers’ output. Though, this explanation does not apply to *Bud* because supervisors do not change the degree of differentiation in the second part.

Furthermore, both the recommendation to differentiate and the financial budget may negatively affect workers’ perceived autonomy. When a recommendation is given, the perceived risk of receiving a poor rating becomes more salient than in the baseline treatment. Similarly, the financial budget raises awareness for negative ratings, and makes good ratings less likely to achieve. Accordingly, both rating policies may impose a cognitive constraint on workers, diminishing their perceived autonomy.⁵⁴ According to self-determination theory, a reduced autonomy lowers motivation (Ryan & Deci, 2000). This reduction in perceived autonomy may become more present over time, potentially explaining why workers decrease their performance in the second part.

The concept of “gift exchange” suggests that higher payoffs may be rewarded with a higher performance driven by positive reciprocity (Fehr et al., 1997). Since ratings are more lenient and payoffs higher in the baseline treatment, positive reciprocity may be stronger than in *Rec* or *Bud*, potentially explaining why performance declines in those treatments in the second part. In *Rec*, the higher differentiation of ratings may lose its motivational effect in the second part. This could be attributed to a shift in workers’ expectations (e.g., low performers no longer accept a bad grade and instead respond with negative reciprocity).

Moreover, fatigue may have contributed to the adjusting performance of *Rec* and *Bud* in the second part. While the rating policies enhance performance in the first part, sustaining this level of effort may have been challenging for subjects due to fatigue. They may have overexerted themselves initially under the rating policies. Though, please note that the performance level is similar under *Base* in the second part of the experiment which serves as a control treatment.

7. Conclusion

We propose two rating policies that can reduce supervisors’ subjective biases and (initially) increase subsequent worker performance. Firstly, we show that supervisors consider the recommendation to give differentiated ratings in their evaluation decisions. Hence, a simple recommendation is an effective method to obtain less lenient ratings and a higher degree of differentiation in performance appraisals. The recommendation provides a simple and cost-effective alternative to FDS, making it suitable for organizations seeking practical solutions to oppose rating compression and leniency. Secondly, the financial budget inevitably leads to less lenient appraisals, but the results are ambiguous regarding the degree of differentiation. Additionally, organizations often strive for fairness in income distributions and thus, differentiated appraisals, i.e., employees shall be rewarded according to their actual performance. Both rating policies help to improve the consideration of relative performance standards.

Our results imply that both rating policies are good mechanisms to increase workers’ performance, at least in the short term. When comparing the two, the financial budget is more cost efficient due to a better cost-benefit ratio compared to the recommendation to differentiate. Nevertheless, the cost efficiency should not be the only criterion when selecting a rating policy. Supervisors feel significantly more uncomfortable to evaluate the workers in *Bud* compared to *Base* whereas feeling uncomfortable is not a concern for the supervisors in *Rec*. This discomfort signals the supervisors’ reluctance to give bad evaluations in *Bud*. Indeed, about 80% of supervisors in *Bud* states they would have given different evaluations with a higher budget. In contrast, approximately 70% of workers believe that a higher budget would not have influenced supervisors’ grading decisions, suggesting that workers do not perceive the budget constraint as decisive for the grading. Thus, we observe a mismatch between supervisors’ and workers’ perceptions.

In both *Bud* and *Rec*, workers’ opinions of the supervisors are worse than in *Base* whereby workers seem to be more dissatisfied in *Bud*. This suggests that workers are more satisfied with evaluations when no specific rating policy is implemented. Additionally, workers perceive the evaluations in *Base* as fairer compared to *Bud*. In the long run, supervisors’ and workers’ satisfaction with a rating policy may impact its success. Evidence from the post-experimental questionnaire supports this assumption. In *Base*, workers hardly change their mind regarding the statements between the first and second part of the experiment. However, in *Bud* and *Rec*, workers significantly change their opinions on nearly half of all statements. This shift suggests reduced satisfaction and motivation in the second part, which may explain the decline in performance in the second part.

While we infer practical implications from our results, they have to be viewed with caution as real organizational settings differ from our laboratory experiment. In our experiment, supervisors and workers were anonymous, whereas in organizations, workers typically know their supervisors and colleagues. In an organizational setting, supervisors may consider their personal relationship with

⁵⁴ For example, Abeler et al., (2011) find that individuals constrained by a reference point tend to exhibit lower performance. Similarly, Marchegiani et al., (2016) observe that individuals constrained by more lenient or severe ratings demonstrate a reduced performance.

the subordinates when evaluating performance (Bol, 2011; Breuer et al., 2013). Consequently, supervisors may hesitate to give bad evaluations, potentially reducing the effectiveness of the recommendation to differentiate in an organizational setting.⁵⁵

Some companies may face restricted monetary resources forcing them to set a financial budget. In this situation, companies may consider whether it would be helpful to recommend differentiated performance appraisals. In *Bud*, our findings show that differentiated ratings lead to a decreasing performance. Therefore, a recommendation to give differentiated ratings may worsen this effect, making the combination of a budget and a recommendation potentially counterproductive. Future research is needed to investigate this question further.

To conclude, both a simple recommendation and a financial budget are effective short-term policies for increasing workers' performance. However, neither rating policy sustains an enhanced performance over time, as both policies may undermine perceived autonomy and fairness, thereby reducing workers' satisfaction and motivation.

From a short term perspective, the recommendation is particularly appealing due to its simplicity, low costs, and ability to mitigate compressed and lenient ratings. While both policies enhance performance levels initially, the mechanisms driving these increases appear to differ between the two treatments. Consistent with prior research (Bol, 2011; Engellandt & Riphahn, 2011; Kampkötter & Sliwka, 2018), our findings demonstrate that higher differentiation in appraisals fosters higher performance if it is recommended to differentiate. Though, if there is a restricted budget for performance ratings, we observe that allocating similar ratings instead of differentiated ratings enhances performance. Thus, other factors such as positive reciprocity or perceptions of fairness in connection with compressed ratings may play a more substantial role in driving the increased performance observed under the budget restriction.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Dirk Sliwka and Johannes Berger for valuable input, and Johannes Berger for supporting data collection. Additionally, we would like to thank the participants of the 2022 *World ESA Meeting* at the MIT Sloan School of Management in Cambridge and the 2022 *European ESA Meeting* at University of Bologna for their valuable feedback.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.jebo.2025.107394](https://doi.org/10.1016/j.jebo.2025.107394).

Appendix

Table A1

Summary descriptive statistics of all treatments.

	Periods	Base (n = 64)				Rec (n = 64)				Bud (n = 64)			
		Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Preliminary Period Group		45.94	13.96	20.50	72.00	45.97	13.92	18.50	69.50	49.13	13.24	23.50	73.50
Output													
Group Output	1–8	61.90	12.80	36.50	99.00	66.08	13.03	32.50	97.00	69.65	15.53	32.00	108.50
	9–16	68.13	12.77	41.50	101.50	67.96	16.20	29.50	107.00	69.09	19.03	18.00	121.50
	1–16	65.02	13.15	36.50	101.50	67.02	14.72	29.50	107.00	69.34	17.36	18.00	121.50
Group Timeout	1–8	0.60	1.61	0.00	8.00	0.41	1.14	0.00	7.00	0.83	1.98	0.00	7.00
	9–16	0.58	1.07	0.00	8.00	1.48	2.76	0.00	10.00	2.41	3.75	0.00	14.00

(continued on next page)

⁵⁵ Other factors may also influence subjective performance evaluations. For example, Reiter et al., (2024) show that virtual backgrounds in digital work settings may bias subjective evaluations.

⁵⁶ When controlling for workers' personality traits and the propensity to be positive or negative reciprocal, the main effects remain robust.

⁵⁷ When controlling for workers' personality traits and the propensity to be positive or negative reciprocal, the main effects remain robust.

Table A1 (continued)

	Periods	Base (n = 64)				Rec (n = 64)				Bud (n = 64)			
		Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Individual Grade	1–16	0.59	1.37	0.00	8.00	0.94	2.18	0.00	10.00	1.62	3.01	0.00	14.00
	1–8	1.67	0.98	1.00	5.00	2.10	1.09	1.00	5.00	3.07	0.98	1.00	5.00
	9–16	1.64	0.91	1.00	5.00	2.21	1.23	1.00	5.00	3.06	1.11	1.00	5.00
Span of Grades	1–16	1.65	0.95	1.00	5.00	2.15	1.16	1.00	5.00	3.06	1.05	1.00	5.00
	1–8	1.32	1.13	0.00	4.00	1.78	0.94	0.00	4.00	1.79	1.16	0.00	4.00
	9–16	1.13	1.01	0.00	4.00	1.95	1.14	0.00	4.00	2.15	1.32	0.00	4.00
	1–16	1.23	1.07	0.00	4.00	1.86	1.05	0.00	4.00	1.97	1.26	0.00	4.00

Table A2

Percentage of grade distribution according to treatment.

Grade Distribution	Base (n = 256)		Rec (n = 256)		Bud (n = 256)	
	(1)	(2)	(3)	(4)	(5)	(6)
(1,1,1)	70	27.34 %	18	7.03 %	/	/
(2,2,2)	3	1.17 %	2	0.78 %	/	/
(3,3,3)	0	0.00 %	3	1.17 %	53	20.70 %
(4,4,4)	0	0.00 %	1	0.39 %	0	0.00 %
(5,5,5)	0	0.00 %	2	0.78 %	3	1.17 %
(1,2,2)	30	11.72 %	15	5.86 %	/	/
(1,3,3)	6	2.34 %	3	1.17 %	/	/
(1,4,4)	0	0.00 %	0	0.00 %	11	4.30 %
(1,5,5)	0	0.00 %	0	0.00 %	4	1.56 %
(2,1,1)	53	20.70 %	29	11.33 %	/	/
(2,3,3)	3	1.17 %	6	2.34 %	/	/
(2,4,4)	0	0.00 %	2	0.78 %	2	0.78 %
(2,5,5)	0	0.00 %	0	0.00 %	1	0.39 %
(3,1,1)	16	6.25 %	21	8.2 %	/	/
(3,2,2)	5	1.95 %	3	1.17 %	/	/
(3,4,4)	1	0.39 %	2	0.78 %	1	0.39 %
(3,5,5)	0	0.00 %	0	0.00 %	0	0.00 %
(4,1,1)	5	1.95 %	1	0.39 %	/	/
(4,2,2)	3	1.17 %	1	0.39 %	/	/
(4,3,3)	1	0.39 %	4	1.56 %	7	2.73 %
(4,5,5)	0	0.00 %	0	0.00 %	0	0.00 %
(5,1,1)	5	1.95 %	1	0.39 %	/	/
(5,2,2)	0	0.00 %	1	0.39 %	38	14.84 %
(5,3,3)	1	0.39 %	0	0.00 %	0	0.00 %
(5,4,4)	0	0.00 %	0	0.00 %	0	0.00 %
(1,2,3)	28	10.94 %	72	28.13 %	/	/
(1,2,4)	10	3.91 %	14	5.47 %	/	/
(1,2,5)	2	0.78 %	14	5.47 %	/	/
(1,3,4)	3	1.17 %	4	1.56 %	/	/
(1,3,5)	3	1.17 %	9	3.52 %	25	9.77 %
(1,4,5)	1	0.39 %	0	0.00 %	1	0.39 %
(2,3,4)	6	2.34 %	18	7.03 %	107	41.8 %
(2,3,5)	1	0.39 %	7	2.73 %	1	0.39 %
(2,4,5)	0	0.00 %	1	0.39 %	1	0.39 %
(3,4,5)	0	0.00 %	2	0.78 %	1	0.39 %

Columns (1), (3) and (5) show absolute numbers, and columns (2), (4) and (6) show percentages. The slashes in column (5) and (6) signal that these distributions do not exist in *Bud* due to the budget constraint.

Table A3

Impact share of output on share of bonus.

Dependent Variable:	Share Bonus				
	(1) Base	(2) Rec	(3) Bud	(4) Base vs. Rec	(5) Base vs. Bud
Share Output	1.004*** (0.135)	1.110*** (0.104)	1.105*** (0.080)	0.987*** (0.135)	0.984*** (0.135)
Share Output × Rec				0.119 (0.172)	
Rec				−0.040 (0.057)	
Share Output × Bud					0.120 (0.156)
Bud					−0.040

(continued on next page)

Table A3 (continued)

Dependent Variable:	Share Bonus				
	(1) <i>Base</i>	(2) <i>Rec</i>	(3) <i>Bud</i>	(4) <i>Base vs. Rec</i>	(5) <i>Base vs. Bud</i>
Constant	−0.001 (0.045)	−0.037 (0.035)	−0.035 (0.027)	0.004 (0.045)	(0.052) 0.005 (0.045)
Observations	768	762	759	1530	1527
Number of Subjects	48	48	48	96	96
Wald Chi2	-	-	-	-	-

Random Effects Regression (period dummies included). Reference category: *Base*.

Columns (1), (2) and (3): Observations are based on 48 workers and 16 periods.

Columns (4) and (5): Observations are based on 48 workers, 16 periods and 2 treatments.

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A4

Impact budget on span of grades.

Dependent Variable:	Span of Grades			
	(1) Part 1	(2) Part 1	(3) Part 2	(4) Part 2
<i>Bud</i>	0.559** (0.221)	0.692*** (0.259)	0.638*** (0.202)	0.801*** (0.296)
SD of Output	0.139*** (0.024)	0.137*** (0.025)	0.160*** (0.018)	0.164*** (0.021)
Group Output	−0.019** (0.008)	−0.018** (0.008)	−0.016*** (0.005)	−0.016*** (0.005)
Male		0.053 (0.213)		0.058 (0.222)
Age		0.043 (0.060)		0.029 (0.045)
Economics		−0.090 (0.187)		−0.146 (0.228)
Selfish		0.313 (0.343)		0.237 (0.213)
Envy		0.202 (0.346)		0.573* (0.340)
Equity		−0.072 (0.245)		0.191 (0.265)
Constant	1.717*** (0.435)	0.557 (1.495)	1.446*** (0.392)	0.591 (1.323)
Observations	256	248	256	248
Number of Subjects	32	31	32	31
Wald Chi2	86.14	128.57	414.93	375.48

Random Effects Regression (period dummies included). Reference category: *Base* and *Altruism*.

Observations are based on 16 supervisors, 2 treatments and 8 periods.

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A5

Experimental questionnaire supervisors.

Item	<i>Base</i>				<i>Rec</i>				<i>Bud</i>				<i>Base vs. Rec</i>				<i>Base vs. Bud</i>			
	Mean		WSR test		Mean		WSR test		Mean		WSR test		MWU test		MWU test		MWU test		MWU test	
	P1	P2			P1	P2			P1	P2			P1	P2			P1	P2		
			z	p			z	p			z	p			z	p			z	p
I think that the workers put a lot of effort into the task.	6.19	5.94	1.19	0.36	5.31	4.81	1.71	0.13	5.75	4.69	2.58	0.01	1.88	0.06	1.52	0.13	1.24	0.23	2.24	0.02
As a supervisor, I consider myself capable of making appropriate evaluations.	6.13	6.38	−2.00	0.13	5.75	5.81	−0.07	1.00	5.94	6.06	−1.41	0.50	0.30	0.78	0.92	0.41	0.55	0.55	1.17	0.28

(continued on next page)

Table A5 (continued)

Item	Base				Rec				Bud				Base vs. Rec				Base vs. Bud			
	Mean		WSR test		Mean		WSR test		Mean		WSR test		MWU test		MWU test		MWU test		MWU test	
	P1	P2			P1	P2			P1	P2			P1	P2			P1	P2		
	z		p		z		p		z		p		z		p		z		p	
My behavior towards the workers was fair.	6.56	6.25	0.07	1.00	6.06	6.13	-0.05	1.00	6.38	6.13	0.65	0.50	-0.09	0.87	-0.42	0.68	0.30	0.85	0.67	0.54
In my evaluations, I judged the abilities of the workers accurately.	5.44	5.13	0.55	0.61	5.06	5.19	-1.00	1.00	5.44	5.19	0.65	0.50	0.57	0.59	-0.23	0.85	-0.43	0.66	-0.23	0.82
It would have been fairer, or it was the fairest, to give all workers the same payment regardless of their performance.	2.50	3.06	-1.73	0.25	2.25	2.31	-1.00	1.00	2.88	2.88	0.00	1.00	0.58	0.58	1.19	0.24	-0.56	0.59	0.33	0.75
It was important to me that not one of the three workers earned much less than the others in a round.	3.25	3.13	0.05	1.00	3.38	3.56	-1.41	0.50	3.69	3.63	1.00	1.00	-0.18	0.87	-0.64	0.53	-0.67	0.54	-0.67	0.54
I gave negative evaluations to punish the workers.	3.81	3.38	1.41	0.50	3.06	3.50	-1.73	0.25	4.06	5.31	-2.43	0.03	0.97	0.34	-0.17	0.87	-0.36	0.73	-2.42	0.01
I made use of the whole range of grades.	2.88	2.25	1.73	0.25	3.00	3.00	0.00	1.00	5.38	5.50	-1.00	1.00	-0.22	0.83	-0.72	0.49	-2.76	0.00	-3.47	0.00
The relative comparison of the workers' performance played an important role in my evaluations.	6.06	6.06	-0.90	0.63	5.88	5.75	0.62	0.75	6.69	6.56	1.00	1.00	0.50	0.62	0.77	0.46	-2.27	0.02	-1.63	0.13
It was uncomfortable for me to evaluate the workers.	2.31	2.50	-0.05	1.00	2.06	1.75	1.41	0.50	3.94	3.94	0.00	1.00	0.22	0.84	0.69	0.49	-1.89	0.06	-1.81	0.08

For within treatment comparisons the Wilcoxon Signed-Rank test (WSR test) is applied and for between treatment comparisons the Mann-Whitney U test (MWU test).

Table A6

Experimental questionnaire workers.

Item	Base				Rec				Bud				Base vs. Rec				Base vs. Bud			
	Mean		WSR test		Mean		WSR test		Mean		WSR test		MWU test		MWU test		MWU test		MWU test	
	P1	P2			P1	P2			P1	P2			P1	P2			P1	P2		
	z		p		z		p		z		p		z		p		z		p	
I put a lot of effort into the task.	6.02	5.79	3.51	0.00	5.60	4.96	3.92	0.00	5.77	5.06	3.41	0.00	0.81	0.42	2.31	0.02	1.03	0.30	2.44	0.01
I really wanted to get the best	5.81	5.85	-0.97	0.33	5.88	5.58	1.91	0.06	5.98	5.35	3.45	0.00	-0.80	0.42	0.36	0.72	-0.60	0.55	1.32	0.19

(continued on next page)

Table A6 (continued)

Item	Base				Rec				Bud				Base vs. Rec				Base vs. Bud			
	Mean		WSR test		Mean		WSR test		Mean		WSR test		MWU test		MWU test		MWU test		MWU test	
	P1	P2			P1	P2			P1	P2			P1	P2			P1	P2		
			z	p			z	p			z	p	z		p		z		p	
evaluation in my group.																				
It would have been fairer to give all workers the same payment regardless of their performance.	3.19	3.21	-1.00	0.32	3.08	3.15	-1.73	0.08	3.13	3.21	-1.15	0.25	0.11	0.91	0.02	0.99	0.05	0.96	-0.14	0.89
When I got a positive evaluation, it motivated me to put in more effort in the next round.	5.21	5.10	1.01	0.31	5.25	4.83	2.56	0.01	5.25	4.85	2.58	0.01	-0.41	0.68	0.39	0.70	0.35	0.73	0.94	0.35
When I got a negative evaluation, it motivated me to put in more effort in the next round.	4.81	4.81	0.00	1.00	4.94	4.50	2.02	0.04	4.44	3.92	2.62	0.01	-0.38	0.71	0.49	0.63	0.96	0.34	2.03	0.04
My supervisor evaluated me negatively in order to punish me.	3.85	3.79	0.51	0.61	3.77	4.10	-2.36	0.02	4.04	4.10	-0.03	0.98	0.18	0.86	-0.72	0.47	-0.34	0.73	-0.76	0.45
I consider my supervisor capable of making appropriate evaluations.	4.65	4.65	0.27	0.79	4.38	3.96	2.53	0.01	4.35	4.00	0.48	0.63	0.89	0.37	1.70	0.09	0.98	0.33	1.82	0.07
My supervisor's behavior towards the workers was fair.	4.90	4.90	0.01	0.99	4.63	4.29	1.45	0.15	4.31	4.10	0.61	0.54	0.53	0.60	1.19	0.23	1.67	0.09	2.08	0.04
My supervisor judged my abilities accurately in his evaluations.	4.73	4.67	1.24	0.21	4.29	3.90	1.95	0.05	3.88	3.56	1.69	0.09	1.21	0.23	2.05	0.04	2.29	0.20	2.90	0.00
I would feel comfortable letting my supervisor control my payoff in the future as well.	4.08	3.98	1.00	0.32	3.71	3.63	0.03	0.98	3.25	3.06	1.43	0.15	0.73	0.46	0.70	0.49	1.78	0.08	2.03	0.04
I have the impression that the intensity of my efforts played an important role in my evaluations.	4.88	4.98	-0.47	0.64	4.46	4.15	2.64	0.01	4.08	3.88	1.44	0.15	1.05	0.30	1.97	0.05	1.76	0.08	2.66	0.01

For within treatment comparisons the Wilcoxon Signed-Rank test (WSR test) is applied and for between treatment comparisons the Mann-Whitney U test (MWU test).

Table A7
Impact of the recommendation to differentiate on output with controls.⁵⁶

Dependent Variable:	Output		Log Output	
	<i>Base vs. Rec</i> (periods 1 – 8)		<i>Base vs. Rec</i> (periods 1 – 8)	
	(1)	(2)	(3)	(4)
<i>Rec</i>	1.386** (0.588)	1.789** (0.727)	0.054** (0.0269)	0.069** (0.032)
Individual Ability	0.604*** (0.065)	0.568*** (0.072)	0.030*** (0.003)	0.028*** (0.003)
Male		0.901 (0.665)		0.053* (0.028)
Age		−0.158** (0.070)		−0.007*** (0.002)
Economics		−0.084 (0.813)		−0.025 (0.038)
Fun		0.705*** (0.199)		0.025*** (0.008)
Constant	7.514*** (1.002)	8.489*** (1.980)	2.339*** (0.048)	2.424*** (0.086)
Observations	768	768	758	758
Number of Subjects	96	96	96	96
Wald Chi ²	668.24	1111.19	608.45	1388.04

Random Effects Regression (Period dummies included). Reference category: *Base* & Female.

Observations are based on 48 workers, 2 treatments and 8 periods (in (3) and (4), 10 workers with an output of zero are excluded since the logarithm is undefined for zero). Robust standard errors in parentheses (clustered on group_id) *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A8
Impact of budget on output with controls.⁵⁷

Dependent Variable:	Output		Log Output	
	<i>Base vs. Bud</i> (periods 1 – 8)		<i>Base vs. Bud</i> (periods 1 – 8)	
	(1)	(2)	(3)	(4)
<i>Bud</i>	1.856** (0.766)	1.582** (0.698)	0.077** (0.032)	0.064** (0.030)
Individual Ability	0.683*** (0.076)	0.658*** (0.086)	0.032*** (0.003)	0.031*** (0.003)
Male		0.626 (0.715)		0.032 (0.027)
Age		−0.201** (0.084)		−0.008*** (0.002)
Economics		−0.393 (0.900)		−0.030 (0.038)
Fun		0.883*** (0.235)		0.028*** (0.009)
Constant	6.038*** (1.252)	7.539*** (2.601)	2.297*** (0.051)	2.392*** (0.092)
Observations	768	768	752	752
Number of Subjects	96	96	96	96
Wald Chi ²	403.09	459.24	481.67	692.66

Random Effects Regression (Period dummies included). Reference category: *Base* & Female.

Observations are based on 48 workers, 2 treatments and 8 periods (in (3) and (4), 16 workers with an output of zero are excluded since the logarithm is undefined for zero). Robust standard errors in parentheses (clustered on group_id) *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A9

Impact of low, middle and top performer on individual timeouts.

Dependent Variable:	Individual Timeouts	
	<i>Rec</i> Periods 1–16 (1)	<i>Bud</i> Periods 1–16 (2)
Middle Performer	0.070 (0.059)	0.234 (0.146)
Low Performer	0.509*** (0.191)	1.374*** (0.364)
Constant	0.010 (0.100)	−0.329 (0.207)
Observations	768	768
Number of Subjects	48	48
Wald Chi ²	-	-

Random Effects Regression (period dummies included). Reference category: Top Performer.

Observations are based on 48 workers and 16 periods.

Robust standard errors in parentheses (clustered on group_id). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.**Table A10**Impact of previous grade on individual timeouts in *Rec*.

Dependent Variable:	Individual Timeouts _t					
	(1) All Part 1 & 2	(2) All Part 1	(3) All Part 2	(4) Low/Middle Part 1 & 2	(5) Low/Middle Part 1	(6) Low/Middle Part 2
Grade2 _{t-1}	0.057 (0.072)	−0.127* (0.069)	0.222* (0.122)	0.121 (0.149)	−0.214* (0.125)	0.343 (0.224)
Grade3 _{t-1}	0.082 (0.103)	−0.058 (0.121)	0.242 (0.181)	0.104 (0.165)	−0.110 (0.188)	0.315 (0.270)
Grade4 _{t-1}	0.727*** (0.233)	0.124 (0.185)	1.304*** (0.441)	0.797*** (0.252)	0.089 (0.239)	1.424*** (0.475)
Grade5 _{t-1}	2.488** (1.236)	−0.151 (0.097)	2.050 (1.281)	2.982** (1.247)	−0.272* (0.155)	3.079** (1.453)
Constant	0.208*** (0.075)	0.208*** (0.075)	−0.041 (0.100)	0.269*** (0.100)	0.267*** (0.097)	−0.209 (0.175)
Observations	768	384	384	491	241	250
Number of Subjects	48	48	48	48	46	48
Wald Chi ²	-	26.40	140.31	-	35.96	118.50

Random Effects Regression (period dummies included). Reference category: Grade1_{t-1}. Observations are derived from workers and period combinations. Robust standard errors in parentheses (clustered on group_id).*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.**Table A11**Impact of previous grade on individual timeouts in *Bud*.

Dependent Variable:	Individual Timeouts _t					
	(1) All Part 1 & 2	(2) All Part 1	(3) All Part 2	(4) Low/Middle Part 1 & 2	(5) Low/Middle Part 1	(6) Low/Middle Part 2
Grade1 _{t-1}	−0.283** (0.117)	−0.096* (0.058)	−0.203 (0.192)	−0.045 (0.404)	0.066 (0.111)	0.003 (0.498)
Grade2 _{t-1}	−0.120 (0.118)	−0.0003 (0.084)	−0.098 (0.180)	−0.098 (0.164)	−0.022 (0.152)	0.016 (0.357)
Grade4 _{t-1}	0.443* (0.229)	0.343* (0.179)	0.457 (0.316)	0.515* (0.267)	0.453** (0.225)	0.496 (0.364)
Grade5 _{t-1}	1.508** (0.637)	0.749 (0.888)	1.865*** (0.562)	1.557** (0.649)	1.325 (1.238)	2.099*** (0.570)
Constant	0.250** (0.109)	0.250** (0.109)	0.491 (0.318)	0.386** (0.164)	0.355** (0.141)	0.700 (0.497)
Observations	768	384	384	490	244	246
Number of Subjects	48	48	48	47	46	46
Wald Chi ²	-	36.00	51.32	-	32.57	37.76

Random Effects Regression (period dummies included). Reference category: Grade3_{t-1}. Grade3_{t-1} functions as reference group because this group is sufficiently large to make inferences about middle/low performers. Observations are derived from workers and period combinations. Robust standard errors in parentheses (clustered on group_id). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A12Impact span of grades on individual subsequent output in *Rec* (periods 1–8).

Dependent Variable:	Individual Output _{t+1}			
	(1) All	(2) Top Performer	(3) Middle Performer	(4) Low Performer
Span of Grades=1 _t	1.862 (1.312)	0.162 (1.432)	3.048 (2.431)	3.492* (2.117)
Span of Grades=2 _t	1.230 (1.433)	−0.300 (1.309)	1.660 (2.443)	3.796 (2.649)
Span of Grades=3 _t	2.668* (1.440)	−0.777 (1.259)	2.911 (2.673)	7.570*** (2.325)
Span of Grades=4 _t	3.947** (1.794)	−4.450 (3.178)	5.822* (3.252)	10.970*** (2.771)
Individual Output _t	0.635*** (0.058)	0.405*** (0.107)	0.547*** (0.128)	0.807*** (0.268)
SD of Output _t	−0.171 (0.123)	0.098 (0.198)	0.172** (0.087)	−0.426 (0.436)
Individual Ability	0.208*** (0.052)	0.414*** (0.085)	0.173* (0.100)	0.067 (0.193)
Constant	4.317** (1.968)	6.321** (2.720)	5.183* (3.026)	2.210 (3.294)
Observations	384	143	107	134
Number of Subjects	48	40	42	38
Wald Chi ²	6723.18	1177.01	272.86	1088.67

Random Effects Regression (period dummies included). Reference category: Span of Grades = 0_t. Observations are derived from workers and period combinations. Robust standard errors in parentheses (clustered on group_id). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A13Impact span of grades on individual subsequent output in *Bud* (periods 1–8).

Dependent Variable:	Individual Output _{t+1}			
	(1) All	(2) Top Performer	(3) Middle Performer	(4) Low Performer
Span of Grades=1 _t	0.013 (2.547)	4.062 (2.540)	1.669 (2.274)	−7.327 (6.582)
Span of Grades=2 _t	−1.742 (1.173)	−0.642 (1.539)	−0.892 (1.010)	−3.225** (1.543)
Span of Grades=3 _t	−2.668** (1.263)	1.937* (1.161)	−4.323 (3.238)	−6.368** (2.542)
Span of Grades=4 _t	−3.684** (1.467)	0.405 (1.929)	0.092 (2.143)	−11.450*** (4.109)
Individual Output _t	0.481*** (0.102)	0.558*** (0.102)	0.498*** (0.139)	0.035 (0.298)
SD of Output _t	0.443*** (0.169)	0.233** (0.117)	0.270 (0.183)	0.079 (0.434)
Individual Ability	0.425*** (0.116)	0.463*** (0.147)	0.396*** (0.121)	0.552 (0.353)
Constant	6.216*** (2.264)	4.294* (2.361)	5.887*** (2.184)	13.410*** (4.951)
Observations	384	140	111	133
Number of Subjects	48	38	43	38
Wald Chi ²	475.57	2368.38	3008.69	279.20

Random Effects Regression (period dummies included). Reference category: Span of Grades = 0_t. Observations are derived from workers and period combinations. Robust standard errors in parentheses (clustered on group_id). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

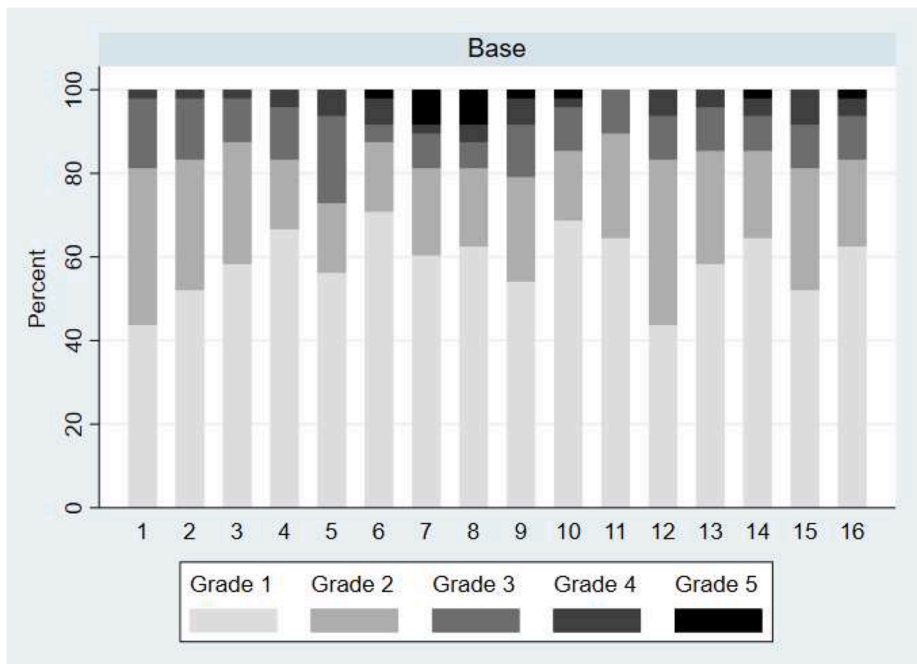


Fig. A1. Grade distribution per period in *Base* on individual level.

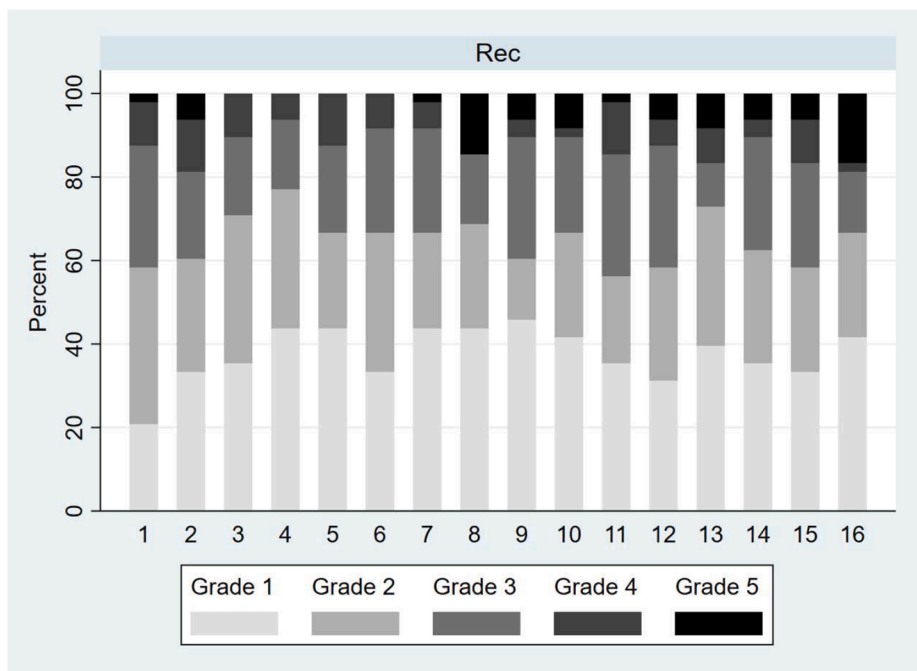


Fig. A2. Grade distribution per period in *Rec* on individual level.

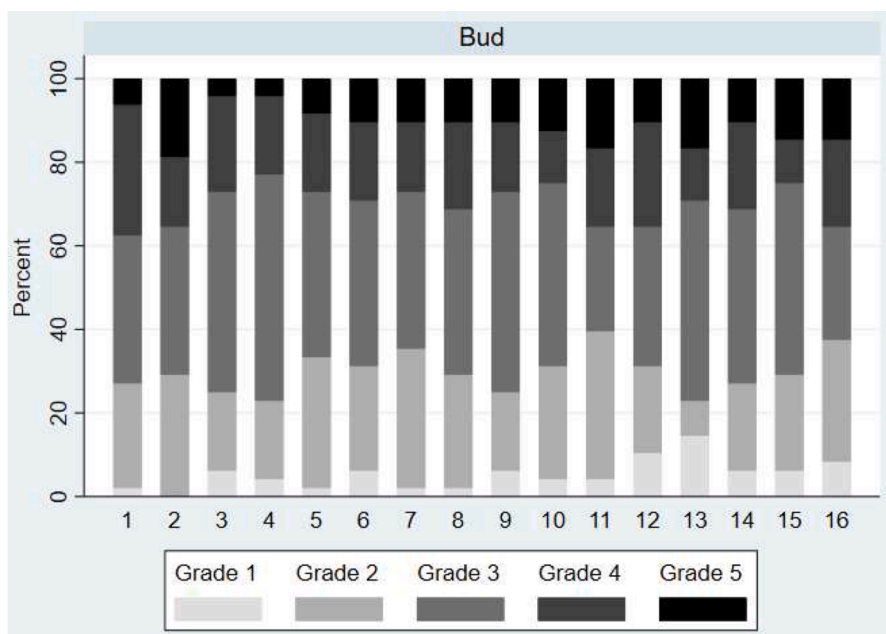


Fig. A3. Grade distribution per period in *Bud* on individual level.

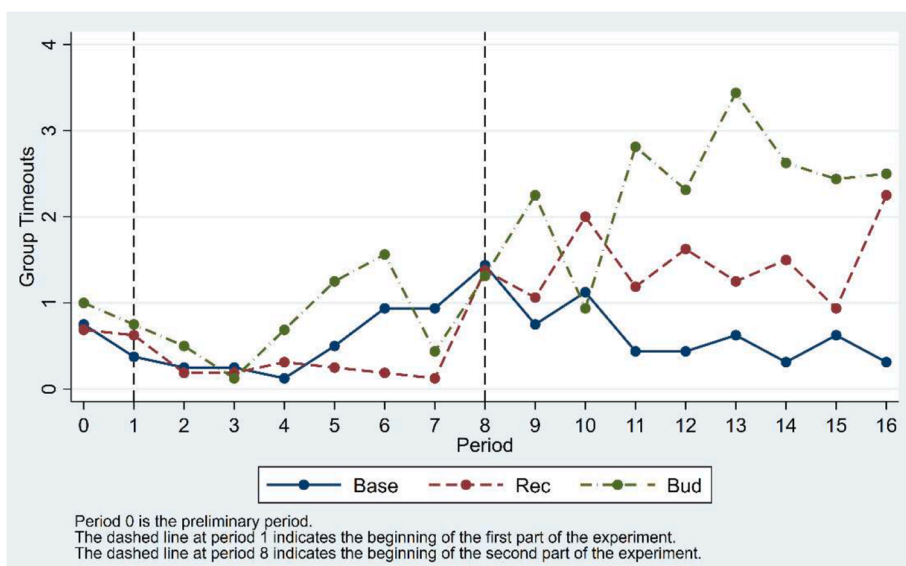


Fig. A4. Group timeouts periods 0–16 on group level.

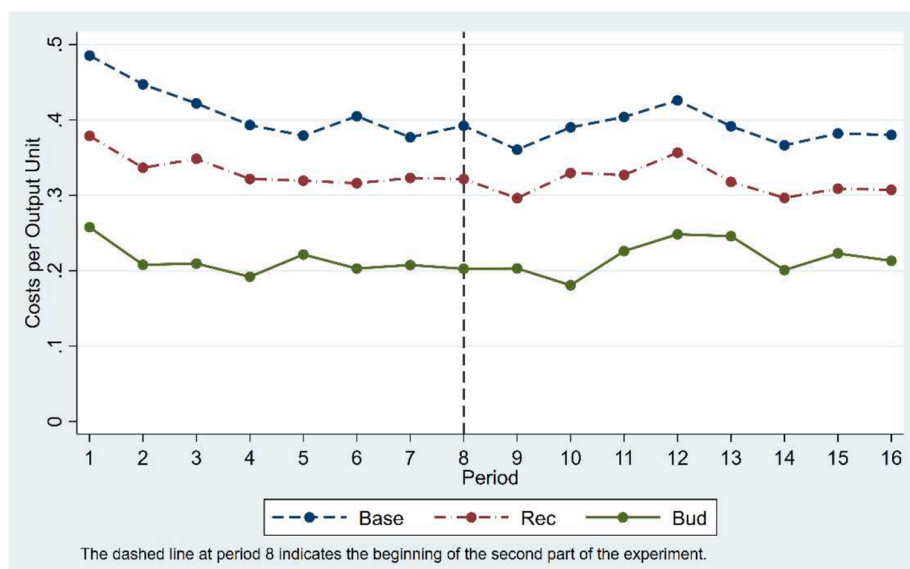


Fig. A5. Costs per output unit periods 1–16 on individual level.

Data availability

Data will be made available on request.

References

- Abeler, J., Falk, A., Goette, L., Huffman, D., 2011. Reference points and effort provision. *Am. Econ. Rev.* 101 (2), 470–492. <https://doi.org/10.1257/aer.101.2.470>.
- Ahn, T.S., Hwang, I., & Kim, M.-I. (2008, November 3). Discriminability of subjective performance measures and its impact on ratee incentives. <https://doi.org/10.2139/ssrn.1302954>.
- Benson, A.M., Sajjadi, S., 2018. Are bonus pools driven by their incentive effects? Evidence from fluctuations in gainsharing incentives. *ILR Rev.* 71 (3), 567–599. <https://doi.org/10.1177/0019793917726066>.
- Berger, J., Harbring, C., Sliwka, D., 2013. Performance appraisals and the impact of forced distribution—An experimental investigation. *Manage Sci.* 59 (1), 54–68. <https://doi.org/10.1287/mnsc.1120.1624>.
- Bernardin, H.J., Cooke, D.K., Villanova, P., 2000. Conscientiousness and agreeableness as predictors of rating leniency. *J. Appl. Psychol.* 85 (2), 232–236. <https://doi.org/10.1037/0021-9010.85.2.232>.
- Bernardin, H.J., Thomason, S., Ronald Buckley, M., Kane, J.S., 2016. Rater rating-level bias and accuracy in performance appraisals: the impact of rater personality, performance management competence, and rater accountability. *Hum. Resour. Manage* 55 (2), 321–340. <https://doi.org/10.1002/hrm.21678>.
- Blanco, M., Engelmann, D., Normann, H.T., 2011. A within-subject analysis of other-regarding preferences. *Games. Econ. Behav.* 72 (2), 321–338. <https://doi.org/10.1016/j.geb.2010.09.008>.
- Bol, J.C., 2011. The determinants and performance effects of managers' performance evaluation biases. *Account. Rev.* 86 (5), 1549–1575. <https://doi.org/10.2308/accr-10099>.
- Brandts, J., Macleod, W.B., 1995. Equilibrium selection in experimental games with recommended play. *Games Econ. Behav.* 11 (1), 36–63. <https://doi.org/10.1006/game.1995.1040>.
- Breuer, K., Nieken, P., Sliwka, D., 2013. Social ties and subjective performance evaluations: an empirical investigation. *Rev. Manag. Sci.* 7 (2), 141–157. <https://doi.org/10.1007/s11846-011-0076-3>.
- Bruns, H., Perino, G., 2021. Point at, nudge, or push private provisions to a public good? Field experimental evidence for experts, politicians, and nobodies. *Econ. Inq.* 1–12. <https://doi.org/10.2139/ssrn.3178597>.
- Charness, G., 2004. Attribution and reciprocity in an experimental labor market. *J. Labor. Econ.* 22 (3), 665–688. <https://doi.org/10.1086/383111>.
- Cleveland, J.N., Murphy, K.R., Williams, R.E., 1989. Multiple uses of performance appraisal: prevalence and correlates. *J. Appl. Psychol.* 74 (1), 130–135. <https://doi.org/10.1037/0021-9010.74.1.130>.
- Cohen-Charash, Y., Spector, P.E., 2001. The role of justice in organizations: a meta-analysis. *Organ. Behav. Hum. Decis. Process.* 86 (2), 278–321. <https://doi.org/10.1006/obhd.2001.2958>.
- Croson, R., Marks, M., 2001. The effect of recommended contributions in the voluntary provision of public goods. *Econ. Inq.* 39 (2), 238–249. <https://doi.org/10.1111/j.1465-7295.2001.tb00063.x>.
- Dannenberger, A., Riechmann, T., Sturm, B., & Vogt, C. (2007). Inequity aversion and individual behavior in public good games: an experimental investigation (ZEW Discussion Paper 07-034).
- Dohmen, T., Falk, A., Huffman, D., Sunde, U., 2009. Homo reciprocans: survey evidence on behavioural outcomes. *Econ. J.* 119 (536), 592–612. <https://doi.org/10.1111/j.1468-0297.2008.02242.x>.
- Eisenhardt, K.M., 1989. Agency theory: an assessment and review. *Acad. Manag. Rev.* 14 (1), 57–74. <https://doi.org/10.5465/amr.1989.4279003>.
- Engellandt, A., Riphahn, R.T., 2011. Evidence on incentive effects of subjective performance evaluations. *ILR Rev.* 64 (2), 241–257. <https://doi.org/10.1177/001979391106400202>.
- Fehr, E., Gächter, S., Kirchsteiger, G., 1997. Reciprocity as a contract enforcement device: experimental evidence. *Econometrica* 65 (4), 833–860. <https://www.jstor.org/stable/2171941>.
- Fischbacher, U., 2007. Z-Tree: zurich toolbox for ready-made economic experiments. *Exp. Econ.* 10 (2), 171–178. <https://doi.org/10.1007/s10683-006-9159-4>.

- Greiner, B., 2004. The Online Recruitment System ORSEE - A guide For the Organization of Experiments in Economics. Max Planck Institute of Economics.
- Grund, C., Przemek, J., 2012. Subjective performance appraisal and inequality aversion. *Appl. Econ.* 44 (17), 2149–2155. <https://doi.org/10.1080/00036846.2011.560109>.
- Harbring, C., Irlenbusch, B., 2011. Sabotage in tournaments: evidence from a laboratory experiment. *Manage Sci.* 57 (4), 611–627. <https://doi.org/10.1287/mnsc.1100.1296>.
- Jawahar, I.M., Williams, C.R., 1997. Where all the children are above average: the performance appraisal purpose effect. *Pers. Psychol.* 50 (4), 905–925. <https://doi.org/10.1111/j.1744-6570.1997.tb01487.x>.
- Kahneman, D., Miller, D.T., 1986. Norm theory: comparing reality to its alternatives. *Psychol. Rev.* 93 (2), 136–153. <https://doi.org/10.1037/0033-295x.93.2.136>.
- Kampkötter, P., Sliwka, D., 2016. The complementary use of experiments and field data to evaluate management practices: the case of subjective performance evaluations. *J. Institutional Theor. Econ.* 172 (2), 364–389. <https://doi.org/10.1628/093245616x14545727832367>.
- Kampkötter, P., Sliwka, D., 2018. More dispersion, higher bonuses? On differentiation in subjective performance evaluations. *J. Labor. Econ.* 36 (2), 511–549. <https://doi.org/10.1086/694588>.
- Kneeland, N., 1929. That lenient tendency in rating. *Pers. J.* 7, 356–366. <https://psycnet.apa.org/record/1929-02810-001>.
- Krueger, J., Mueller, R.A., 2002. Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *J. Pers. Soc. Psychol.* 82 (2), 180–188. <https://doi.org/10.1037/0022-3514.82.2.180>.
- Kube, S., Maréchal, M.A., Puppe, C., 2012. The currency of reciprocity: gift exchange in the workplace. *Am. Econ. Rev.* 102 (4), 1644–1662. <https://doi.org/10.1257/aer.102.4.1644>.
- Landy, F.J., Farr, J.L., 1980. Performance rating. *Psychol. Bull.* 87 (1), 72–107. <https://doi.org/10.1037/0033-2909.87.1.72>.
- Lazear, E.P., Rosen, S., 1981. Rank-order tournaments as optimum labor contracts. *J. Polit. Econ.* 89 (5), 841–864. <https://doi.org/10.1086/261010>.
- Loberg, L., Nüesch, S., Foege, J.N., 2021. Forced distribution rating systems and team collaboration. *J. Econ. Behav. Organ.* 188, 18–35. <https://doi.org/10.1016/j.jebo.2021.04.035>.
- Marchegiani, L., Reggiani, T., Rizzolli, M., 2016. Loss averse agents and lenient supervisors in performance appraisal. *J. Econ. Behav. Organ.* 131, 183–197. <https://doi.org/10.1016/j.jebo.2016.07.022>.
- Meyer, H.H., 1975. The pay-for-performance dilemma. *Organ. Dyn.* 3 (3), 39–50. [https://doi.org/10.1016/0090-2616\(75\)90029-7](https://doi.org/10.1016/0090-2616(75)90029-7).
- Moers, F., 2005. Discretion and bias in performance evaluation: the impact of diversity and subjectivity. *Account. Organ. Soc.* 30 (1), 67–80. <https://doi.org/10.1016/j.aos.2003.11.001>.
- Mohnen, A., Pokorny, K., Sliwka, D., 2008. Transparency, inequity aversion, and the dynamics of peer pressure in teams: theory and evidence. *J. Labor. Econ.* 26 (4), 693–720. <https://doi.org/10.1086/591116>.
- Mursita, L.Y., Nahartyo, E., 2022. How centrality bias in subjective evaluation affects positive and negative employee work behavior: a real-effort task experiment. *J. Account. Organ. Change* 18 (5), 789–810. <https://doi.org/10.1108/JAOC-01-2020-0002>.
- Prendergast, C., 1999. The provision of incentives in firms. *J. Econ. Lit.* 37 (1), 7–63. <https://doi.org/10.1257/jel.37.1.7>.
- Rammstedt, B., John, O.P., 2007. Measuring personality in one minute or less: a 10-item short version of the big five inventory in english and german. *J. Res. Pers.* 41 (1), 203–212. <https://doi.org/10.1016/j.jrp.2006.02.001>.
- Reiter, K., Mohnen, A., Walkowitz, G., 2024. The impact of virtual computer screen backgrounds on subjective performance evaluation in online videoconferences. *Comput. Hum. Behav. Rep.* 16, 100516. <https://doi.org/10.1016/j.chbr.2024.100516>.
- Ryan, R.M., Deci, E.L., 2000. Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25 (1), 54–67. <https://doi.org/10.1006/ceps.1999.1020>.
- Trapp, I., Trapp, R., 2019. The psychological effects of centrality bias: an experimental analysis. *J. Bus. Econ.* 89 (2), 155–189. <https://doi.org/10.1007/s11573-018-0908-6>.