Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Performance Appraisals and the Impact of Forced Distribution—An Experimental Investigation

Johannes Berger, Christine Harbring, Dirk Sliwka,

To cite this article:
Johannes Berger, Christine Harbring, Dirk Sliwka, (2013) Performance Appraisals and the Impact of Forced Distribution—An Experimental Investigation. Management Science 59(1):54-68. http://dx.doi.org/10.1287/mnsc.1120.1624

MANAGEMENT SCIENCE
Vol. 59, No. 1, January 2013, pp. 54–68
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1624

Performance Appraisals and the Impact of Forced Distribution—An Experimental Investigation

Johannes Berger
University of Cologne, 50931 Cologne, Germany, johannes.berger@uni-koeln.de

Christine Harbring
RWTH Aachen University, 52056 Aachen, Germany; and IZA–Institute for the Study of Labor, 53113 Bonn, Germany, christine.harbring@rwth-aachen.de

Dirk Sliwka
University of Cologne, 50931 Cologne, Germany; IZA–Institute for the Study of Labor, 53113 Bonn, Germany; and CESifo, 81679 Munich, Germany, dirk.sliwka@uni-koeln.de

A real-effort experiment is investigated in which supervisors have to rate the performance of individual workers who in turn receive a bonus payment based on these ratings. We compare a baseline treatment in which supervisors are not restricted in their rating behavior to a forced distribution system in which they have to assign differentiated grades. We find that productivity is significantly higher under a forced distribution by about 6% to 12%. However, the productivity effects are less clear cut when participants have prior experience with the baseline condition. Moreover, a forced distribution becomes detrimental when workers have access to a simple option to sabotage each other.

Key words: performance measurement; forced distribution; forced ranking; motivation; experiment
History: Received June 17, 2010; accepted March 23, 2012, by Peter Wakker, decision analysis. Published online in *Articles in Advance* October 8, 2012.

1. Introduction

In most jobs an employee's true efforts are at best imprecisely captured by objective figures. Hence, organizations frequently use subjective appraisals to evaluate substantial parts of an employee's job performance. Whereas this may strengthen the setting of incentives as more facets of job performance are evaluated, the opposite may be true when supervisors bias evaluations according to personal preferences.[1]

There is indeed strong evidence from numerous studies indicating that subjective performance ratings tend to be biased. First of all, it has often been stressed that supervisors are too "lenient" and reluctant to use the lower spectrum of possible performance ratings. Moreover, supervisors typically do not differentiate enough between high and low performers such that ratings tend to be compressed relative to the distribution of the true performance outcomes.[2] Because rating scales nearly always have an upper boundary, rater leniency often directly implies rating compression. Whereas the existence of these biases has been confirmed in previous studies, the mechanisms behind these biases and the effects on performance have only rarely been analyzed empirically. Rynes et al. (2005, p. 572), for instance, stressed that "although there is a voluminous psychological literature on performance evaluation, surprisingly little of this research examines the consequences of linking pay to evaluated performance in work settings."

A simple economic logic suggests that both of the above mentioned biases should lead to weaker incentives when rewards are tied to ratings. When high performance is not rewarded and low performance is not sanctioned adequately, employees should have lower incentives to exert effort when they anticipate biased ratings. In contrast, it may be argued that rating leniency can trigger positive reciprocity, and rating compression reduces inequity among coworkers, which both may lead to increased employee motivation.[3]

To avoid potential negative consequences of rater biases, some firms have adopted so-called "forced

[1] For an overview, see, for instance, Murphy and Cleveland (1995) and Arvey and Murphy (1998), or from an economics perspective, see Prendergast and Topel (1993, 1996) or Gibbs et al. (2003).

[2] These two biases are often referred to in the literature as the "leniency" bias and "centrality" bias. See, for instance, Landy and Farr (1980), Murphy (1992), Bretz et al. (1992), Jawahar and Williams (1997), Prendergast (1999), and Moers (2005).

[3] Many experimental studies have now confirmed that higher wage payments indeed trigger positive reciprocity and in turn can lead to higher efforts. See, for instance, Fehr et al. (1993, 1997), Hannan et al. (2002), or Charness (2004). Evidence from field experiments is somewhat less pronounced. Recent studies found mostly moderate support for positive reciprocity. See, for instance, Gneezy and List (2006), Cohn et al. (2009), Kube et al. (2012), Bellemare and Shearer (2009), and Hennig-Schmidt et al. (2010).

distribution" systems under which supervisors have to follow a predetermined distribution of ratings. At General Electric, for example, the former chief executive officer Jack Welch promoted what he called a "vitality curve" according to which each supervisor had to identify the top 20% and the bottom 10% of his team in each year. In practice, the use of such systems remains controversially discussed (Holland 2006).

From an economic perspective, forced distribution systems have the structure of rank-order tournaments (see Lazear and Rosen 1981), in which contestants compete for a limited number of prizes. In a forced distribution system workers compete for one of the scarce good performance ratings that are typically associated with a monetary reward, e.g., a bonus or a salary increase. A well-known downside of tournaments, however, is the danger that cooperation among workers is put at risk because there is always an incentive to improve one's relative position by increasing one's productive effort but also by harming others, i.e., sabotaging others (Lazear 1989). A similar argument could hold for forced rankings as well.

A key reason for the lack of field evidence on the consequences of a forced distribution is that even when a firm changes its performance appraisal system, there is typically no control group within the same firm with an unaltered scheme. This in turn makes it hard to identify the causal effect of the modification. Moreover, to measure performance consequences objective performance measures are necessary. But such measures are typically not available when subjective assessments are used.

In this paper we investigate the performance consequences of a forced distribution system in a real-effort experiment. In each experimental group, one participant in the role of a supervisor has to evaluate the performance of three participants in the role of employees over several rounds. Participants have to work on a real-effort task where the outcome of their work directly determines the supervisor's payoff. At the end of each round the supervisor learns the work outcome of each individual employee and is asked to individually rate their performance on a five-point scale. The employees then receive a bonus payment based on this performance rating. We examine two experimental settings: In the baseline treatment, supervisors are not restricted in their rating behavior. In a forced distribution treatment, they have to give differentiated ratings. We also investigate additional treatments in which a forced distribution system is either abolished or introduced at a later date, a setting in which supervisors share the bonus costs paid out to subordinates, and another setting in which subordinates can sabotage each other.

Our key result is that when there is no possibility to interfere with the colleagues' work, productivity in our experiment is approximately 6%–12% higher under a forced distribution system. Moreover, we find that in the absence of a forced distribution system, supervisors who care more for the well-being of others tend to assign more lenient and therefore less differentiated ratings. But weaker degrees of differentiation lead to lower performance in subsequent rounds. Interestingly, supervisors seem to learn the advantages of differentiation as they assign less lenient and more differentiated ratings after the forced distribution has been abolished compared to a setting in which it has never been used. On the other hand, the performance effect of a forced distribution is less clear cut when participants have experienced the more "liberal" baseline setting before and, hence, have different reference standards and expectations.

The forced distribution is also beneficial in a treatment in which it is costly for the supervisors to assign high bonuses. But interestingly, there is now an additional reason for the treatment difference: The forced distribution protects high-performing workers from "stingy bosses" who are unwilling to pay high bonuses. Our key result is, however, reversed when we study additional treatments in which employees can sabotage their colleagues' work. In this setting, the between-worker competition induced by a forced distribution generates detrimental effects that outweigh potential productivity gains.

Although to the best of our knowledge there are no previous studies investigating the effects of the introduction of a forced distribution on incentives, some recent field studies investigate the effects of rating compression on future outcomes. Engellandt and Riphahn (2011), Bol (2011), and Kampkötter and Sliwka (2011) gave some indication that rating compression is associated with lower subsequent performance. Direct empirical evidence on the effects of forced distributions is very scarce. Recently, Schleicher et al. (2009) experimentally investigated raters' reactions to forced distribution and found that rating decisions are perceived as more difficult and less fair under a forced distribution system than in a traditional setting. Scullen et al. (2005) conducted a simulation study and showed that forced distribution can increase performance in the short run as low performers are driven out of the firm, but this effect becomes smaller over time. Neither study examines the incentive effects of forced distributions.

2. Experimental Design

We conduct a laboratory study investigating several different treatments. In all treatments, subjects in the role of "supervisor" evaluate the performance of other subjects in the role of "workers" who have to work on a real-effort task. Supervisors benefit from

higher work efforts in all treatments. For each setting we compare a baseline treatment in which supervisors are not restricted in their evaluation behavior to a forced distribution treatment. Each treatment consists of several parts, which are described in the following.

2.1. Ability Test

In an initial preround, all subjects have to work on the real-effort task that is also used in the main part of the experiment, i.e., all participants have to repeatedly count the number "7" in blocks of randomly generated numbers. This preround is conducted to collect a measure for each subject's ability for the task and to familiarize participants with the task (also those who are in the role of the supervisor). We have chosen the particular design of the task for several reasons: First, the task is tedious and requires real work effort. Second, work outcomes are observable for supervisors and the experimenter, i.e., we have a precise measure of performance that can be compared between the otherwise identical treatments. Third, noise does not play a substantial role for performance. And finally, it is possible to assess the subjects' ability and give the supervisors some experience with the task before the experiment.

To ensure everybody has correctly understood the task, an "exercise block" is presented on the computer screen prior to the preround. Only after all subjects have correctly solved this block, the preround, which lasts 2.5 minutes, is started. During the preround, each subject's performance is measured by the number of collected "points." For each correct answer, a subject receives two points, and for each wrong answer, 0.5 points are substracted. At the end of the preround a piece-rate of 10 (euro) cents per point is paid to each participant's account. During the task, subjects can use a "time-out" button that locks their screen for 20 seconds, during which subjects cannot work on any blocks. Each time the time-out button is pushed, the subject receives 8 (euro) cents, representing potential further opportunity cost of working. At the end of the preround, each participant is informed about the total number of points as well as the number of correct and false answers and the resulting payoff.[4]

2.2. Main Part: Performance Ratings and Bonus Payments

After the ability test, instructions for the first part of the experiment are distributed. Before this part of the experiment starts, participants have to answer several test questions on the screen to making sure they have fully understood the procedures and the payoff calculations. This first part of the experiment consists of eight periods, each lasting for 2.5 minutes. Each participant is assigned to a group consisting of four participants. One participant in each group has the role of the "supervisor," and the other three participants are "workers." The group composition as well as the roles remain fixed throughout the experiment. The workers have to perform the same real-effort task as in the preround. They can again make use of a time-out button, blocking the screen for 20 seconds, for which they receive 25 (euro) cents on their private account. After each round, each worker learns his total number of points, the number of correct and false answers, and the number of time-outs chosen. Moreover, each worker is also informed about the number of points and correct and false answers of the other two workers in his group. The supervisor also receives this individual performance information for each of the three workers in her group and then has to rate each worker on a rating scale of 1 to 5, with 1 being the best and 5 the worst rating available.

Each rating is associated with a bonus payment for the worker, ranging from 10€ for the highest rating, 1, to 0€ for the worst rating, 5, in steps of 2.5€. It is important to stress that in our core setting the supervisor does not personally bear the costs of the bonus payments. The reason is that in most field settings supervisors are not residual claimants but are themselves salaried employees. Higher bonus payments to subordinates, therefore, do no lower their own income. In an extension we also investigate treatments in which supervisors bear costs for higher bonus payments.

The round payoff for the worker is the sum of his bonus payment and the payoff from pushing the time-out button. The payoff of the supervisor is solely determined by the output of the three workers in her group. For each point achieved by one of the three workers the supervisor receives 30 (euro) cents. At the end of the round, each worker is informed about his rating, the number of time-outs, and his resulting payoff. The worker does not learn about the other workers' ratings in his group. In each part of the experiment, one round is randomly selected to be payoff relevant (for details, see §2.5).

2.3. Matching of Groups

To avoid a situation in which performance ratings are predetermined by substantial ability differences, we match participants into homogeneous groups. The matching procedure is based on the performance in the preround; i.e., all 32 subjects are individually ranked in each session based on their total number of points achieved in the preround. The four participants with the best ranking are assigned to a group, the four best individuals of the remaining participants to the next group, etc. Within each group, the participant with the best performance is assigned the role

[4] To avoid losses, the total number of points for a period were set to zero when the total for this period was negative.

of the supervisor. Participants are not informed about the matching procedure to avoid strategic considerations. Subjects only know they will be grouped with three other participants. At the end of the experiment, a few additional decision games are played, and a short questionnaire is filled out.

2.4. Treatments

In our core setting we analyze two different treatments: In the baseline treatment (*Base*) supervisors are not restricted in their rating behavior. In the forced distribution setting (*Fds*) supervisors have to give one worker a rating of 1 or 2, one worker a rating of 3, and another worker a rating of 4 or 5. This restriction is explained to all participants in the treatment.

To analyze the effects of introducing or abolishing a forced distribution system in a within-subject design, we split the experiment into two parts, each consisting of eight consecutive rounds. The group matching as well as the assigned roles are kept constant across both parts. In our treatment *BaseFds*, for example, participants work in the baseline setting for eight rounds (first part), which are followed by eight rounds of the forced distribution setting (second part). To disentangle rating rule effects from time and learning effects, we conduct two additional treatments in which the rating rule does not change across both parts of the experiment (*BaseBase* and *FdsFds*).

2.5. Procedures

After participants arrived in the laboratory, they were seated in separated cabins where they received the instructions for the preround of the experiment. Participants were told not to communicate. In case of any question they had to raise their hand such that one of the experimenters would come and help. The experiment started after all participants had read the instructions and all questions had been answered. After the preround, instructions for the first part of the experiment were distributed. Instructions for the second part only followed after the first part had been completed.

The instructions informed participants that only one of the eight rounds of each part of the experiment was payoff relevant for all participants. At the end of each session, a randomly selected subject was asked to twice draw one of eight cards to determine which rounds were to be paid out. The final payoff for each subject consisted of the money earned during the experiment and a show-up fee of 4€. The money was paid out anonymously in cash at the end of each session.

In total, the core setting of the experiment consisted of eight sessions with two sessions for each treatment condition. Thus, we have 64 subjects (16 independent groups) in each treatment with a total of 256 participants. It was ensured that no one had been involved in an experiment with the same real-effort task before. No subject participated in more than one session. On average, a session lasted for 2.5 hours, and the average payoff amounted to 27€. The experiment was conducted at the Cologne Laboratory for Economic Research. All sessions were computerized using the experimental software *z*-Tree (Fischbacher 2007), and subjects were recruited with the online recruiting system ORSEE (Greiner 2004).

3. Results

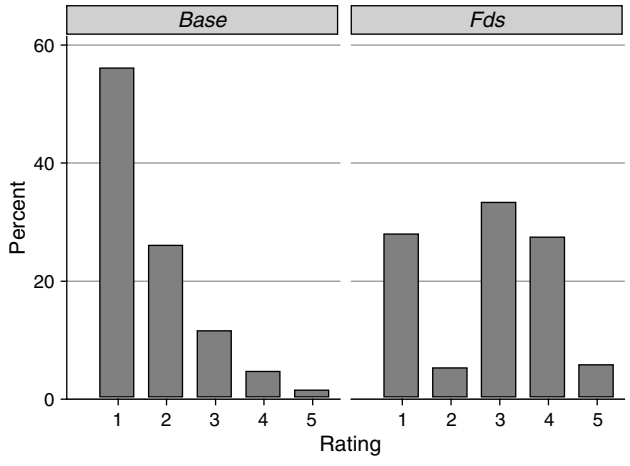
In this section we first give an overview of the performance effect of the forced distribution system in our core experimental setting. We then analyze the driving forces behind the observed treatment differences in more detail. Section 3.4 provides an overview of spillover effects observed when the sequence of both settings varies in *BaseFds* and *FdsBase*. Finally, we report the results of two additional experiments, one in which awarding bonuses is costly for the supervisors and one in which subordinates can sabotage each other.

3.1. Performance Effects of Forced Distribution

We start with an analysis of the first part. For each of the two treatment conditions (*Fds* and *Base*) we thus have 32 strictly independent group observations.

Figure 1 contrasts the distribution of ratings in the first eight periods in *Base* and *Fds*. Evidently, supervisors in *Base* tend to assign very good ratings, i.e., a 1 or 2, in the majority of cases (82%). Note that this pattern closely resembles the typical "leniency bias" often observed in organizational practice. Bretz et al. (1992, p. 333), for instance, described this as follows: "even though most organizations report systems with five levels, generally only three levels are used. . . . It is common for 60%–70% of an organization's workforce to be rated in the top two performance levels.

Figure 1 Distribution of Ratings Across Treatments



. . . Skewed performance distributions not only exist, but are common." As in most real-world organizations, supervisors in the experiment do not have to bear the direct costs of higher bonus payments. In this situation they indeed have a tendency to assign high bonuses to their subordinates, a behavior limited by the forced distribution system. Nonetheless, within the degrees of freedom left by the system, the supervisors in *Fds* still follow the lenient choices and strongly prefer the rating 1 over 2 and the rating 4 over 5, as shown in the right panel of Figure 1.

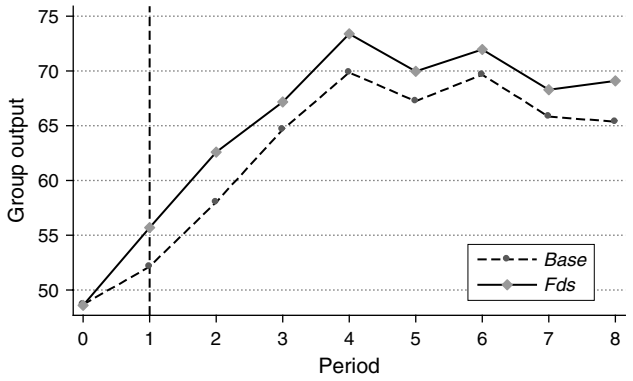
But it is of course important to investigate the performance consequences of this rating behavior. A key hypothesis based on a simple economic reasoning is that the return to effort should be lower in the baseline treatment compared to the forced distribution treatment. Hence, workers should have lower incentives to exert high effort levels. Instead, one may argue that supervisors assign good grades on purpose, hoping to trigger positive reciprocity on the workers' part and thereby increasing their motivation.

Figure 2 plots the average group output over time and shows that performance indeed increases under the forced distribution. Although performance is identical prior to the treatment intervention (in the preround), average performance is higher under the forced distribution across all periods of the experiment.[5]

We investigate the size and significance of the performance effect by running three different regression specifications with either group output (the sum of individual outputs per group) or individual output as the dependent variable. Because of the matching procedure, we control for the number of points achieved by the group or the individual in the preround (period 0).

As a first, conservative econometric approach that preserves the independence of observations, we compute the group average over all eight periods and regress it on a treatment dummy and the preround performance using only one data point per group.[6] Because the group observations are independent and the treatment intervention exogenous, the estimated coefficient of the forced distribution dummy gives a clean estimate of the average treatment effect. In the second specification we use all group observations over time (i.e., jointly achieved group points per period) and run random effects regressions that include period dummies to control for the general time trend observed in Figure 2. In a third alternative

Figure 2 Distribution of Group Output Over Time Across Treatments



Notes. Period 0 is the preround. The dashed line at period 1 indicates the beginning of the first part of the experiment.

we use observations from all individual workers in all periods again estimating a random effects model. We report standard errors clustered on the group level to account for the fact that observations from workers in the same group are not independent. The results are reported in Table 1. In the left panel we run all three specifications using absolute output measures; in the right panel we report specifications with the log of output as the dependent variable.

Column (1) shows that the forced distribution indeed significantly increases group performance by approximately three output units. This corresponds to a 5.6% increase in group performance as displayed in model (4).[7] The coefficients obtained in the random effects models parallel these results. Furthermore, in all specifications preround performance is strongly correlated with actual performance in the experiment.

Investigating the treatment differences with alternate productivity measures, such as the number of blocks finished per group and the number of correct and false answers (see Table A.2 in the appendix), we find that under forced distribution subjects counted and solved more blocks correctly while making only slightly and insignificantly more mistakes. The time-out option was rarely chosen in the two core treatments (see Table A.1 in the appendix), and there was no systematic difference in time-out usage across treatments.

To provide an even more conservative test of the key hypothesis without any distributional assumptions, we additionally apply the following nonparametric procedure: Because of our matching mechanism, groups within a treatment are by definition not drawn from the same population, but groups with the same preround performance rank are directly comparable across treatments. We thus rank

[5] It is interesting to note that the qualitative shapes of the graphs over time are quite similar, reflecting parallel effects of learning and fatigue.

[6] Similar results are obtained when only using the outputs from period 1.

[7] Note that in log specification the coefficient of 0.054 translates into an estimated incease of 5.6% as $e^{0.054} = 1.056$.

Table 1 Impact of Forced Distribution on Productivity

Dependent variable:	*Output*			Log *output*		
	Base vs. *Fds* (periods 1–8)			*Base* vs. *Fds* (periods 1–8)		
	(1) OLS	(2) RE (groups)	(3) RE (individuals)	(4) OLS	(5) RE (groups)	(6) RE (individuals)
Fds	3.197**	3.197**	1.066**	0.0540**	0.0540**	0.0699***
	(1.562)	(1.551)	(0.514)	(0.0242)	(0.0240)	(0.0235)
Preround group output	0.534***	0.534***	0.530***	0.00842***	0.00842***	0.0251***
	(0.0550)	(0.0546)	(0.0531)	(0.00087)	(0.00086)	(0.00245)
Constant	38.11***	26.35***	8.847***	3.725***	3.529***	2.398***
	(2.701)	(2.540)	(0.834)	(0.0449)	(0.0441)	(0.0443)
Observations	64	512	1,536	64	512	1,509
Number of groups/subjects	64	64	192	64	64	192
R^2/Wald χ^2	0.68	741.10	723.90	0.69	532.75	523.67

Note. Robust standard errors are in parentheses (and in (3) and (6), clustered on *group_id*). In columns (2), (3), (5), and (6), period dummies included. Columns (1) and (4) show ordinary least squares (OLS) regression on average group output (one observation per group), (2) and (5) show random effects (RE) regressed on periodic group output, and (3) and (6) show random effects regressed on periodic individual output.

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

group observations in each treatment according to their preround performance from 1 to 32 and calculate the output difference of each group with its counterpart in the other treatment; e.g., the average group output of the eighth able group in the *Fds* condition is compared with the eighth able group in the *Base* condition. If there were no systematic output differences across treatments, we would expect to see balanced output differences between paired groups. However, in 21 out of 32 output comparisons, output was higher in the *Fds* groups. This difference is statistically significant in a one-sided binominal test ($p = 0.055$). Applying the same test to test for differences in preround performance, we see that in 6 out of 32 comparison groups output was exactly the same, 13 times output was higher in *Fds*, and 13 times output was higher in *Base*. Hence, randomization performed very well such that ability is equally distributed across the two treatment groups (see also Figure 2).

We also investigate whether the incentive effect of forced distribution is stronger among low- or high-talent groups. Table A.3 in the appendix extends our standard regression by an interaction term *Fds* × *Preround group output*. The substantially larger and highly significant *Fds* coefficient and the negative interaction term reveal that forced distribution is particularly effective among low-performing groups.

Finally, we explore the performance effect of *Fds* in the second set of eight periods in the treatments *BaseBase* and *FdsFds*, which allows us to check the persistence of the observed effects. Applying the identical identification strategy as above, we find rather similar results, and the economic significance of the effect gets even stronger: The regression results displayed in Table A.4 in the appendix show that the performance difference between *Fds* and *Base* amounts to 8.8% in the second part. The effect is significant across all regressions, and also when we apply the described nonparametric procedure ($p = 0.038$, one-sided Binomial test).[8]

3.2. Rating Behavior and Productivity

But why do people work harder under the forced distribution? A key conjecture is that under the forced distribution incentives to exert effort are strengthened because supervisors differentiate more according to individual performance. We therefore analyze whether performance is rewarded differently in the two treatments. In principle, supervisors can condition their grading behavior on two dimensions: they can reward absolute and relative performance. We naturally should expect that the relative rank plays a key role under the forced distribution. But even in the baseline treatment supervisors may condition their grading behavior on the employee's relative rank in the group. However, they may do so to a smaller extent because they are not forced to differentiate. Variations in absolute performance may affect the grades in both treatments. To test this we run random effects regressions with the bonus received in a period as dependent and the absolute output and

[8] When considering only the *BaseBase* and *FdsFds* treatments there is a significant difference in the preround outputs, indicating that abilities are not equally distributed across treatments in this smaller sample. But as mentioned, abilities are evenly distributed when we consider the larger number of independent observations.

Table 2 Impact of Rank and Output on Bonus Payments

Dependent variable:	*Individual bonus*			
	Base (periods 1–8)	*Fds* (periods 1–8)	*Base* vs. *Fds* (periods 1–8)	*BaseBase* vs. *FdsFds* (periods 9–16)
	(1)	(2)	(3)	(4)
Output	0.284***	0.0922***	0.258***	0.221***
	(0.0337)	(0.0147)	(0.0291)	(0.0415)
Output × *Fds*			−0.155***	−0.121***
			(0.0294)	(0.0407)
Rank 2	0.705***	2.064***	0.780***	0.761***
	(0.196)	(0.152)	(0.179)	(0.234)
Rank 1	0.926***	5.747***	1.047***	1.078***
	(0.344)	(0.373)	(0.326)	(0.303)
Rank 2 × *Fds*			1.159***	1.605***
			(0.245)	(0.251)
Rank 1 × *Fds*			4.424***	5.206***
			(0.518)	(0.368)
Fds			−1.372**	−2.486**
			(0.567)	(1.063)
Preround group output	−0.132***	−0.0434***	−0.0824***	−0.0539**
	(0.0382)	(0.0129)	(0.0195)	(0.0242)
Constant	4.186***	1.870***	3.780***	3.438***
	(0.721)	(0.272)	(0.567)	(0.952)
Observations	768	768	1,536	768
Number of subjects	96	96	192	96
Wald χ^2	468.10	1,762.93	1,855.37	3,382.97

Notes. Robust standard errors are in parentheses (clustered on *group_id*). A random effects regression (period dummies included) is shown. The reference category is Rank 3.

**$p < 0.05$; ***$p < 0.01$.

relative rank of a worker as independent variables.[9] To illustrate treatment differences we include interaction terms with a dummy variable for the forced distribution treatment.

The results are reported in Table 2. Clearly, bonus payments depend on both absolute performance and the worker's relative rank. However, the interaction terms in column (3) illustrate two effects: First, the association between within-rank variation in output and bonuses is higher in the baseline setting. Second, and more importantly, the link between relative ranks and bonuses is much stronger under the forced distribution. This supports the idea that the forced distribution induces a tournament-like environment in which the competition for rank leads to an overall increase in performance.[10] Supervisors in the baseline setting could have created a similar between-worker competition but were reluctant to sufficiently condition their ratings on relative performance ranks.

Figure 3 Distribution of Ratings According to Relative Performance in the Group

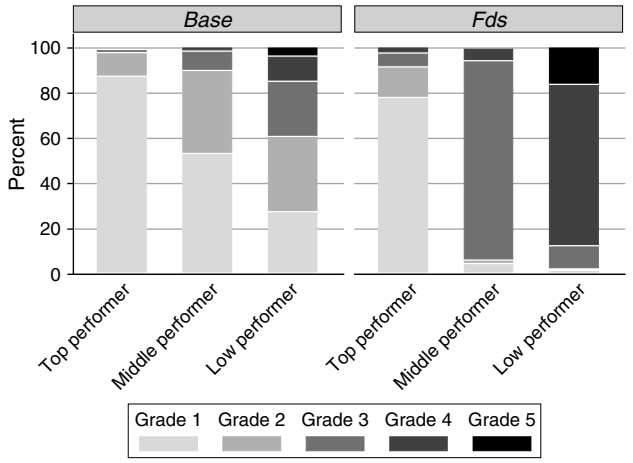


Figure 3 provides a graphical illustration of this result. It shows the distribution of grades for the top, middle, and low performers in the first eight periods across both treatment conditions. In the forced distribution treatment, 91% of the participants with the highest rank receive a 1 or 2, and 88% with the lowest rank receive a 4 or 5. In contrast, approximately 60% of the worst performers still receive a 1 or 2 in the baseline treatment. Clearly, the gains from improving the rank are much weaker in the baseline treatment.

We can also investigate a worker's direct reaction to a particular grade. Table 3 reports results from regressions with individual output in $t+1$ as the dependent variable and dummy variables for the grade assigned in period t as independent variables. Of course selection is an issue here because high performers receive better grades. We thus run random effects regressions controlling for output in period t and preround output of each agent and split up the sample according to relative performance. The reference category corresponds to receiving the top grade, 1. Model (1) analyzes the average reaction of all workers in the baseline setting. Model (2) only includes the observations of the top performers, and model (3) only the observations of the middle and low performers in each period. Because a grade of 5 is rarely observed, we pool grades of 4 and 5.

We observe significant immediate reactions in those cases where the grade obtained is particularly informative about a supervisor's grading behavior: Middle and low performers substantially increase their outputs when receiving a 2 or a 3 compared to receiving the top grade, 1. Thus, those who are not the best performers and yet receive the

[9] The last rank, 3, is the reference group. Note that we do not claim to make clean causal statements because of course the rank is an endogenous variable; rather, we use the regressions to describe patterns of associations.

[10] For experimental evidence on tournaments, see, for example, Schotter and Weigelt (1992), Orrison et al. (2004), and Harbring and Irlenbusch (2011).

Table 3 Impact of Ratings on Individual Performance

Dependent variable:	*Individual output*$_{t+1}$					
	Base (periods 1–8)			*Fds* (periods 1–8)		
	(1) All workers	(2) Top	(3) Middle/low	(4) All workers	(5) Top	(6) Middle/low
Grade $= 2_t$	1.276***	1.074	1.114***	0.480	0.326	−0.348
	(0.330)	(0.761)	(0.425)	(0.540)	(0.744)	(1.657)
Grade $= 3_t$	1.652***	−0.349	1.469**	−1.049*	−2.621	−1.208
	(0.487)	(1.705)	(0.623)	(0.603)	(2.134)	(1.279)
Grade $= 4$ or 5_t	0.977	−0.612	0.587	−1.833**	−8.708***	−1.472
	(1.539)	(2.037)	(1.692)	(0.921)	(2.389)	(1.298)
Output$_t$	0.710***	0.855***	0.628***	0.538***	0.473***	0.494***
	(0.0520)	(0.0707)	(0.0956)	(0.0897)	(0.142)	(0.0963)
Preround output	0.203***	0.193***	0.202***	0.190***	0.246***	0.167**
	(0.0416)	(0.0723)	(0.0681)	(0.0580)	(0.0772)	(0.0718)
Constant	2.852***	0.0356	4.399***	8.713***	9.399***	9.981***
	(0.772)	(1.247)	(1.082)	(1.725)	(3.291)	(1.862)
Observations	672	243	429	672	228	444
Number of subjects	96	79	94	96	72	92
Wald χ^2	1,150.92	1,133.28	585.71	587.55	233.88	180.78

Notes. Robust standard errors are in parentheses (clustered on *group_id*). A random effects regression (period dummies included) is shown. The reference category is Grade $= 1_t$.
$^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

top grade reduce their efforts, which supports the view that lenient and undifferentiated ratings indeed undermine performance incentives.[11] When the forced distribution is in place, subjects know the rating policy because grades are mostly determined by output ranks. In turn, receiving a particular grade does not provide valuable additional information, and indeed we find weaker reactions to grades. However, as can be seen in column (5), top performers on average reduce their efforts after receiving a 4 or a 5. In this case they can directly infer that worse performing coworkers have obtained better grades and that high performance is not rewarded. Time-out choices also depend on grades: Whereas top performers are significantly more likely to take a time-out in response to a 2 instead of a 1, we find that giving worse grades to the middle and low performers reduces the time-outs taken.

These results suggest that less lenient grades in the baseline setting will lead to higher performance. In fact, we do find a positive association between the span of grades in period t (i.e., the difference between the worst and the best rating assigned by the supervisor) and the group output in period $t+1$ in our data. Moreover, the responses to the items[12] "I assigned bad ratings to motivate the workers" and "I assigned bad ratings to sanction the workers" in the postexperimental questionnaire are both positively correlated with group output and the span of grades in the second part of the experiment (significant at the 10% and 5% levels).

3.3. Social Preferences and Rating Behavior?

As research in personnel psychology[13] has already stressed, the personality of the rater has an effect on evaluation behavior. In the language of (behavioral) economics, we should straightforwardly expect that the supervisor's social preferences such as inequity aversion, altruism, or surplus concerns affect rating decisions. To investigate this we elicit subjects' social preferences before final payoffs are communicated in our experiment.

In particular, there are two direct potential explanations for lenient ratings. On the one hand, throughout all treatments supervisors earn more than workers. In turn, supervisors who are inequity averse (compare Fehr and Schmidt (1999) and Bolton and Ockenfels (2000)) may want to reduce this inequity by assigning better grades. On the other hand, as has been stressed, for instance, by Charness and Rabin (2002), many individuals are also motivated by efficiency concerns (i.e., they may strive to maximize the total surplus of all participants to some extent) or are altruistic and, therefore, directly care for the payoffs of others, and should thus assign better grades resulting in higher bonuses.

[11] This is in line with the experimental study by Abeler et al. (2010), who found that efforts are substantially lower in a multiagent gift exchange experiment when principals are forced to pay all agents the same wage.

[12] For all items we used a seven-point scale running from 1 for "does not apply at all" to 7 for "fully applies."

[13] See, for instance, Kane et al. (1995) or Bernardin et al. (2000).

To investigate these drivers we included an adapted version of an incentivized experimental procedure introduced by Blanco et al. (2010) and modified by Dannenberg et al. (2007) at the end of our experiment. It consists of simple choice experiments in which participants have to choose between pairs of payoff tuples, specifying payments to themselves and to some randomly drawn other subject. In the first set of choices participants have to choose between a rather low but equitable payoff tuple (1, 1) and inequitable tuples with higher overall payoffs but entailing a higher payment to the other subject. In the second set of choices subjects have to choose between a combination of a high payoff for themselves and no payoff for the other subject (5, 0) and equitable tuples that give both participants the same payoff but potentially a lower payoff to the decision maker himself. From the choices in these two games we classify supervisors into four different types. Subjects who only maximize their own payoff are classified as *selfish*. Subjects who (i) reduce their own payoff to increase the other's and the overall surplus but (ii) do not reduce joint surplus to avoid disadvantageous inequity are classified as *altruistic*. Subjects who do the opposite, i.e., they do not reduce their own payoff to increase the overall surplus but reduce the joint surplus to avoid disadvantageous inequity, are *envious*. And finally, those who reduce their own payoff to increase the overall surplus but also reduce joint surplus to avoid disadvantageous inequity are characterized as *equity oriented*.

We expect that both the altruistic and the equity-oriented types assign better grades. We do not expect that envious types' rating behaviors differ from those of selfish types because the supervisors are typically better off than workers.

Table 4 reports regression results with the total bonus payments awarded to the group or the span of grades as dependent variables and dummy variables for the different types as independent variables using data from both parts of the core treatments in which either *Base* or *Fds* was played (the reference group is the selfish supervisors). As expected, we observe that the supervisor's type indeed matters: Altruistic types award the highest grades. Compared to the supervisors classified as selfish, they give an additional 4€ of bonus to their group in each round. The coefficient for the equity oriented types is positive but just fails to be significant. However, column (3) shows that equity-oriented types choose significantly more compressed ratings. Because supervisors earn substantially more than workers, envious supervisors do not rate differently than selfish types. We also investigate to what extent the different supervisor types base their rating decisions on the relative rank and absolute output of the agents. Running the same regressions of Table 2 separately for each supervisor type reveals that rank has the highest effect on the bonus paid out by selfish and envious supervisors, but a much weaker effect for altruistic and equity-oriented ones.

Table 4 What Drives Rating Behavior?

Dependent variable:	*Group bonus*		*Span of grades*	
	Base (periods 1–16)	*Fds*	*Base* (periods 1–16)	*Fds*
	(1)	(2)	(3)	(4)
1 if *Envious*	1.882	−0.149	−0.344	0.0726
	(2.081)	(0.533)	(0.383)	(0.186)
1 if *Altruistic*	3.430**	0.369	−0.760***	0.0336
	(1.594)	(0.299)	(0.232)	(0.102)
1 if *Equity*	2.553	0.0215	−0.545*	−0.0226
	(1.782)	(0.358)	(0.289)	(0.123)
Group output	0.230***	0.0427***	−0.0400***	−0.000122
	(0.0355)	(0.00927)	(0.00648)	(0.00270)
SD of output	−0.300***	−0.0392**	0.171***	0.0669***
	(0.0779)	(0.0189)	(0.0173)	(0.00881)
Preround group output	−0.108***	−0.0221***	0.0156***	−0.00261
	(0.0370)	(0.00777)	(0.00547)	(0.00239)
BaseFds	−0.875	0.101	0.0504	−0.0450
	(1.228)	(0.239)	(0.231)	(0.0749)
FdsBase	−2.168*	−0.199	0.461**	0.0554
	(1.225)	(0.304)	(0.228)	(0.0913)
Constant	14.63***	15.35***	2.727***	2.646***
	(2.429)	(0.630)	(0.382)	(0.203)
Observations	504	472	504	472
Number of groups	47	45	47	45
Wald χ^2	261.53	224.25	393.70	297.09

Notes. Robust standard errors are in parentheses. A random effects regression (period dummies included) is shown. The reference category is 1 = *Selfish supervisors*.

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

3.4. Introducing or Abolishing a Forced Distribution?

In this section we take a closer look at within-treatment variations of forced distribution. In a first step, we investigate the effects of introducing a forced distribution in the second part of the experiment after the agents have experienced the baseline condition in the first part. Taking learning effects into account, we compare the performance in the second part of *BaseFds* with the performance in the second part of *BaseBase*.

Given the results of the between-treatment comparison described in the above, we should expect the forced distribution to increase performance in the second part of *BaseFds*. In contrast, a direct comparison reveals that, on average across all periods of the second part, the introduction of a forced distribution does not lead to a higher performance as shown by column (1) of Table 5. Nevertheless, a surprising

Table 5 Effects of the Introduction of a Forced Distribution

Dependent variable:	*Group output*	
	BaseFds vs. *BaseBase* (periods 9–16)	
	(1)	(2)
BaseFds	−0.855	5.372*
	(2.566)	(3.147)
BaseFds × *Period* 10		−1.594
		(3.432)
BaseFds × *Period* 11		−7.844*
		(4.560)
BaseFds × *Period* 12		−8.125***
		(2.930)
BaseFds × *Period* 13		−8.844**
		(3.541)
BaseFds × *Period* 14		−7.438**
		(3.490)
BaseFds × *Period* 15		−7.781*
		(4.253)
BaseFds × *Period* 16		−8.188***
		(3.108)
Preround group output	0.675***	0.675***
	(0.082)	(0.083)
Constant	40.55***	37.44***
	(4.576)	(4.461)
Observations	256	256
Number of subjects	32	32
Wald χ^2	148.70	325.12

Notes. Robust standard errors are in parentheses. A random effects regression (period dummies included) is shown.
*$p < 0.1$; **$p < 0.05$, ***$p < 0.01$.

Table 6 Introducing and Abolishing Forced Distribution

Dependent variable:	*Group output*		*Group time-outs*	
	BaseFds vs. *FdsFds* (periods 9–16)	*BaseBase* vs. *FdsBase*	*BaseFds* vs. *FdsFds* (periods 9–16)	*BaseBase* vs. *FdsBase*
	(1)	(2)	(3)	(4)
BaseFds	−5.763*		0.735***	
	(2.994)		(0.268)	
FdsBase		4.591*		−0.0055
		(2.363)		(0.015)
Preround group output	0.514***	0.644***	0.00212	0.291
	(0.087)	(0.105)	(0.00725)	(0.375)
Constant	56.09***	41.04***	−0.286	0.959
	(5.187)	(5.312)	(0.377)	(0.724)
Observations	256	256	256	256
Number of groups	32	32	32	32
Wald χ^2	318.06	57.95	19.93	11.07

Notes. Robust standard errors are in parentheses. A random effects regression (period dummies included) is shown.
*$p < 0.1$; ***$p < 0.01$.

pattern emerges when we compare the effects per period as shown in column (2). Although performance first increases by about five points in period 9 and stays at this level in period 10, it drops to approximately two to three points below the baseline level in the last six periods. Apparently, participants are initially motivated to work harder under the forced distribution because they immediately seem to understand that they have to put in higher efforts. However, they quickly learn that it is much harder to attain good grades. In contrast to a setting in which a forced distribution is present from the outset, participants now have a different reference standard because they have experienced more favorable ratings in the past. This may lead to a decrease in motivation under *Fds*, which would be in line with recent field studies by Ockenfels et al. (2010) and Clark et al. (2010) showing that the violation of reference points for bonus payments can have detrimental effects on subsequent performance.

A different explanation would be that forced distribution leads to a different pattern of exhaustion in the second part of the experiment. To test this we compare the last eight periods of *BaseFds* to the treatment in which the forced distribution has been used throughout the experiment (*FdsFds*). But as column (1) of Table 6 shows, the forced distribution system in the second part performs worse after the baseline setting compared to the situation in which agents work under a forced distribution right from the beginning. As a result, it is indeed the experience of the baseline setting with higher grades and bonuses that leads to a demotivational effect of the forced distribution. The negative perception of this relative payment loss apparently seems to counteract the positive forces of increased differentiation. The highly significant difference in time-outs, displayed in column (3), also supports this explanation.

We can also compare the performance of the baseline condition after the experience of a forced distribution to the treatment in which the baseline condition is kept over both parts of the experiment. The positive coefficient of *FdsBase* in column (2) reveals that groups in which *Fds* has been abolished are approximately 7% more productive than workers in *BaseBase*. Analogously to the above reasoning, workers in *FdsBase* seem to be particularly motivated in the second part because they receive (on average) much better grades than under the previous rating scheme. Relative to the workers who have already received inflated ratings over the first eight rounds (*BaseBase*), the workers in *FdsBase* could feel more inclined to reciprocate this relative increase in bonus payments. Yet, another factor driving this result is that supervisors keep up differentiation even after forced distribution has been abolished. We do find some evidence that supervisors in *FdsBase* tend to differentiate more during the second part than their counterparts in *BaseBase* (for a

given output). For a given output, workers ranked second or third in a group are significantly less likely to receive a 1 and are more likely to receive a 4 or 5 in the second part of *FdsBase* than in *BaseBase*. Also, as indicated by the negative *FdsBase* dummy in column (1) of Table 4, ratings are, on average, lower than under *BaseBase*. The experience with a forced distribution apparently helped to establish a norm of making performance-contingent ratings, which leads to a better performance. As a consequence, performance in the second part of *FdsBase* even comes close to performance in *FdsFds*.

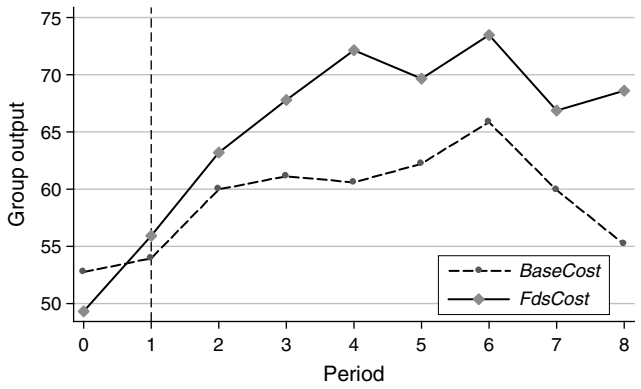
Additional evidence for these arguments comes from our postexperimental questionnaire. We posed to participants who experience both settings in *BaseFds* and *FdsBase* a variety of questions separately for both parts of the experiment. Especially workers in *BaseFds* felt that their effort paid off to a greater extent during the baseline setting. They also stated that the supervisor's behavior was more fair and that she was more capable of giving appropriate ratings in the absence of a forced distribution. The supervisors naturally also expressed some dissatisfaction toward the forced distribution because, for instance, they perceived rating decisions to be more difficult in the second part of *BaseFds*, which is well in line with the findings by Schleicher et al. (2009).

3.5. Forced Distribution and Costly Grades

In most firms the performance of employees is rated by supervisors who themselves are salaried employees. Hence, these supervisors typically do not bear the costs of higher bonus payments. Arguably, they may still have some costs of handing out high bonuses freely. For instance, their own bonus payments may be tied to the compliance with a given bonus budget. Similarly, when a profit sharing scheme is in place, the supervisor's own income is reduced when bonus payments to subordinates are too high. To check the robustness of our results, we therefore investigate a further treatment in which assigning high ratings is costly for the supervisors. In this treatment the supervisor's income is reduced by 50% of the bonus awarded to her agents. To ensure that supervisors always have the possibility to assign the top grade to all of their workers, they are endowed with an additional 15€ per period.[14]

Figure 4 shows the average output over time in the first part of the new treatments.[15] The results are qualitatively surprisingly similar to our earlier results, and the use of a forced distribution here has an even stronger impact on performance.

Figure 4 Distribution of Group Outputs Over Time Across Treatments When Ratings Are Costly



Notes. Period 0 is the preround. The vertical line at period 1 indicates the beginning of the first part of the experiment.

Average group output is 59.6 in *BaseCost* and 67.2 in *FdsCost*. Even though groups in *BaseCost* are on average slightly more productive in the preround, performance is already higher in early periods of *FdsCost*, and this difference increases over time. As the regression results in column (3) of Table A.5 in the appendix show, the performance difference amounts to 9.4 output units, or 12%, and is significant at the 5% level. This difference is also confirmed by the nonparametric testing procedure laid out in the above ($p = 0.059$, one-sided binominal test).

The treatment difference is particularly pronounced in the last periods. This indicates another reason for the superiority of the forced distribution when bonuses are costly. Workers apparently fear that "stingy" supervisors will keep the money for themselves, which is most prevalent toward the end of the experiment. Here the forced distribution protects the bonuses of high-performing workers, which helps to preserve incentives.[16]

3.6. Forced Distribution and Sabotage

The previous sections demonstrated that forcing supervisors to differentiate their evaluations may positively affect performance when workers work on their own. In many jobs, however, workers frequently interact with colleagues and may therefore mutually influence work outcomes. In a positive sense, workers may help and support others to do their work. By the same token, workers may also behave

[14] We added one additional change in this new treatment: Based on the comments of an anoynmous referee, we explicitly told subjects how the supervisor was selected after the preround. We additionally extended the postexperimental questionnaire to check for potential effects of this procedure but did not find evidence that participants' behavior was affected.

[15] One group in *BaseCost* had to be dropped because of a technical problem with the experimental software.

[16] We also studied a second part where participants worked for another eight periods under the same rules. Here the treatment difference is no longer significant in all periods. Although the treatment coefficient is still substantial (6.99), the standard error is now much higher.

uncooperatively, deny help, or even sabotage coworkers. Examples of such behavior could be withholding viable information or, in the extreme, deleting files on computers, stealing others' equipment, or the like. With regard to systems of forced distribution, Prendergast and Topel (1993 p. 362) claimed: "Forced rankings also increase competition for merit pay, which is counterproductive in environments where cooperation is important to production."

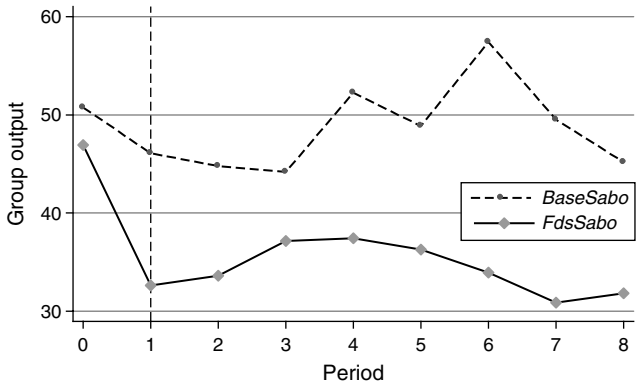
We test this conjecture with a simple treatment variation of our current experimental setup. In addition to counting numbers and taking time-outs, subjects are explicitly given the opportunity to block a coworker's screen for 20 seconds such that the fellow worker cannot work or take time-outs. This "sabotage option" is costly because the choice of blocking somebody else's screen blocks the worker's own screen for three seconds. There is no restriction on the frequency of sabotage, i.e., subjects can block other subjects as often as they like. After being blocked for 20 seconds, it is ensured that subjects can not be sabotaged again within the next five seconds of that period. Sabotage is anonymous, i.e., the sabotaged worker does not know by whom she is sabotaged.

Again, we study this setting over two parts of eight periods, keeping the two treatment conditions baseline (*BaseSabo*) and forced distribution (*FdsSabo*) unchanged in both parts. The key hypothesis is that the forced distribution should lead to higher sabotage activities as workers compete for the high ratings and can improve their position by harming their coworkers. Together with our prior results we therefore conjecture that a trade-off exists because the forced distribution may increase incentives but should also increase wasteful sabotage activities.

Indeed, we find that subjects use the sabotage option twice as often under the forced distribution (about eight times per group and period instead of four) as under the baseline. This difference leads to detrimental consequences for the overall group performance. Under the forced distribution, average group performance is as low as 33.3 and thus 18 points below the baseline treatment with sabotage. The differences in sabotage choices as well as performance are highly significant in regressions as displayed in columns (5) and (6) in Table A.5 in the appendix. Figure 5 displays the performance over time across the two settings and suggests that the performance difference even increases over time. The treatment difference is also robust when we apply the nonparametric test to compare the differences across groups of the same rank with respect to preround performance from both treatments ($p < 0.01$, one-sided binominal test).

Interestingly, higher degrees of differentiation also lead to more sabotage activity within the baseline treatment alone and thus lower performance in subsequent periods: In a random effects regression with the sabotage activity in period $t+1$ as the dependent variable, the span of grades in period t has a substantial positive coefficient controlling for the sabotage activity in period t. More differentiation clearly sets incentives to outperform coworkers, and the easiest way to do this in this setting is to use the sabotage option. The results for the second part of the experiment are very similar. The differences in performance and sabotage become even larger compared to the first part.

Figure 5 Effect of Forced Distribution on Performance Under Sabotage



Notes. Period 0 is the preround. The vertical line at period 1 indicates the beginning of the first part of the experiment.

4. Conclusion

We studied the impact of a forced distribution in a real-effort experiment in which performance is endogenously evaluated by participants. Our key result is that performance is significantly higher under a forced distribution when workers work independently and may not easily harm each other. The reason for this substantial gain in performance is that many supervisors in the baseline setting are very lenient in their rating decisions, and hence, performance incentives are weak.

We also find some evidence that supervisors' social preferences are key drivers of rating behavior: Altruistic supervisors tend to give higher bonuses, whereas equity-oriented supervisors choose significantly less differentiated ratings.

But our results also indicate potential problems of using a forced distribution. First, it may be problematic to set up a forced distribution when employees have experienced a more "liberal" system of performance evaluations before. We find that introducing a forced distribution into an existing appraisal system leads to a short-term performance increase followed by a rather sharp drop in performance. Apparently, workers initially understand that they need to work harder under a forced distribution, but they

are soon demotivated because they cannot attain the bonus levels they have experienced in the past. In contrast, some experience with the forced distribution in the beginning demonstrates to supervisors the benefits of differentiation, because they tend to differentiate more and are able to maintain a higher performance when forced distribution is abolished again. Second, a forced distribution becomes detrimental when we introduce an easily accessible option to sabotage fellow workers. In these treatments the forced distribution leads to a substantially lower performance because the induced competition indeed provides strong incentives to harm coworkers.

Our results have several interesting implications for the design of performance evaluation schemes in practice. First of all, forced distribution systems may foster a high-performing work culture as sometimes conjectured by practitioners. This should particularly be the case when workers work on their own. But the results of the sabotage treatment indicate that firms should be careful in enforcing differentiation when workers can interact and harm each other. It is, of course, important to stress that we introduced an anonymous and rather "easy to use" technology to sabotage coworkers in the experiment. In field settings it is usually much harder to harm a coworker's performance without being detected. Hence, we do not expect equally substantial levels of counterproductive activities in firms in which forced distributions are implemented. Nonetheless, given the strikingly high frequency of participants using the sabotage option in our experiment, firms should be careful in using forced distribution systems in work contexts where mutually harmful counterproductive activities are easily accessible (or where mutual help is strongly beneficial).

The study also indicates that "history matters." When changing the rules of performance evaluations, system designers have to take the employees' as well as supervisors' reference standards and expectations regarding appraisals and bonus payments into account. These reference standards may carry over to the new system and affect the social, economic, and psychological mechanisms at work in the appraisal process. It may be less problematic to use a forced distribution for newly hired employees or for bonuses paid on top of existing compensation packages. But when employees are used to lenient ratings and generous bonuses, setting up a forced distribution may lead to a violation of reference points and trigger negative reciprocity, which may outweigh potential gains from higher powered incentives.

Acknowledgments

The authors thank department editor Peter Wakker, an anonymous associate coeditor, and three reviewers for very helpful comments and suggestions. They also thank Andreas Staffeldt and Timo Vogelsang for great research assistance in programming and conducting the experiments as well as seminar and conference audiences at the 2009 European Workshop on Experimental and Behavioral Economics in Barcelona, the 2009 Gesellschaft für experimentelle Wirtschaftsforschung Meeting in Essen, the 2009 European Association of Labour Economists Annual Conference in Talinn, the 2009 12th Personalökonomisches Kolloquium in Vienna, the 2010 Research Seminar Universitat Autónoma in Barcelona, the 2010 fifth Alhambra Experimental Workshop in Granada, the 2010 Economic Science Association World Meeting in Copenhagen, the 2011 Paris-Cologne Workshop at the University of Paris, and the 2011 CESifo Area Conference on Behavioral Economics. Financial support by the German Research Foundation [Projects HA 4462/1-1, HA 4462/1-2, Research Group Design & Behavior] is gratefully acknowledged.

Appendix

Table A.1 Summary Statistics of all Treatments

Variables:	*Preround group output*	*Group output*		*Group time-out*		*Group rating*		
Periods:		1–8	9–16	1–8	9–16	1–8	9–16	Number of groups
Base (pooled)	48.67	64.11	71.45	0.54	0.71	1.70	1.75	32
Fds (pooled)	48.63	67.28	73.00	0.84	0.60	2.78	2.74	32
BaseBase	45.93	61.90	68.12	0.60	0.58	1.67	1.64	16
BaseFds	51.41	66.31	70.96	0.47	0.97	1.72	2.75	16
FdsFds	48.12	67.73	75.04	0.52	0.23	2.75	2.74	16
FdsBase	49.13	66.83	74.77	1.17	0.85	2.81	1.86	16
BaseCost	52.77	59.87	64.97	2.68	3.06	2.83	2.65	16
FdsCost	49.34	67.22	70.19	1.25	2.26	2.97	2.97	16
BaseSabo	50.75	48.54	54.18	0.77	0.71	2.00	1.84	16
FdsSabo	46.94	34.20	32.41	0.73	0.52	2.81	2.84	16

Table A.2 Performance Effect of Forced Distribution on Different Output Measures

Dependent variable:	*Finished blocks*	*Correct blocks*	*False blocks*	*False/correct blocks*
	Base vs. *Fds* (periods 1–8)			
	(1)	(2)	(3)	(4)
Fds	1.951*	1.700**	0.251	−0.00087
	(1.034)	(0.803)	(0.476)	(0.0145)
Preround group output	0.283***	0.270***	0.0125	−0.00010*
	(0.0434)	(0.0298)	(0.0189)	(0.0006)
Constant	17.62***	14.04***	3.580***	0.209***
	(2.234)	(1.426)	(1.016)	(0.0332)
Observations	512	512	512	512
Number of groups	64	64	64	64
Wald χ^2	626.24	756.05	13.63	19.53

Note. A random effects regression (period dummies included) is shown. Robust standard errors are in parentheses.
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Table A.3 Impact of Forced Distribution Depending on Ability

Dependent variable:	*Group output*	
	Base vs. *Fds* (periods 1–8)	*BaseBase* vs. *FdsFds* (periods 9–16)
	(1)	(2)
Fds	12.83***	21.47***
	(4.964)	(7.255)
Preround group output	0.635***	0.730***
	(0.0610)	(0.0974)
Fds × *Preround group output*	−0.198**	−0.336**
	(0.0941)	(0.136)
Constant	21.47***	37.22***
	(3.043)	(5.295)
Observations	512	256
Number of groups	64	32
Wald χ^2	768.71	186.13

Note. A random effects regression (period dummies included) is shown. Robust standard errors are in parentheses.
**$p < 0.05$; ***$p < 0.01$.

Table A.4 Impact of Forced Distribution on Productivity in the Last Eight Periods

Dependent variable:	*Output*			Log *output*		
	BaseBase vs. *FdsFds* (period 9–16)			*BaseBase* vs. *FdsFds* (period 9–16)		
	(1) OLS	(2) RE (groups)	(3) RE (individuals)	(4) OLS	(5) RE (groups)	(6) RE (individuals)
Fds	5.786**	5.786**	1.931**	0.0839**	0.0839**	0.0947***
	(2.494)	(2.456)	(0.809)	(0.0360)	(0.0354)	(0.0345)
Preround group output	0.516***	0.516***	0.513***	0.00714***	0.00714***	0.0227***
	(0.0927)	(0.0913)	(0.0898)	(0.00134)	(0.00132)	(0.00372)
Constant	44.44***	47.09***	15.74***	3.876***	3.907***	2.769***
	(4.421)	(4.720)	(1.555)	(0.0661)	(0.0714)	(0.0680)
Observations	32	256	768	32	256	764
Number of groups	—	32	96	—	32	96
R^2/Wald χ^2	0.66	179.43	181.33	0.64	192.19	184.69

Notes. Robust standard errors are in parentheses (and in (3) and (6), clustered on *group_id*). Columns (1) and (4) show ordinary least squares (OLS) regression on collapsed average group output (one observation per group). In (2), (3), (5), and (6), period dummies are included. Columns (2) and (5) show random effects (RE) regression on periodic group output, and (3) and (6) show random effects regression on periodic individual output.
**$p < 0.05$; ***$p < 0.01$.

Table A.5 Impact of Forced Distribution on Productivity—Treatment Overview

Dependent variable:	*Group output*					
	BaseBase vs. *FdsFds* (periods 1–8)	*BaseBase* vs. *FdsFds* (periods 9–16)	*BaseCost* vs. *FdsCost* (periods 1–8)	*BaseCost* vs. *FdsCost* (periods 9–16)	*BaseSabo* vs. *FdsSabo* (periods 1–8)	*BaseSabo* vs. *FdsSabo* (periods 9–16)
	(1)	(2)	(3)	(4)	(5)	(6)
Fds	4.728**	5.786**	9.406**	6.994	−13.00***	−20.72***
	(2.155)	(2.456)	(3.978)	(6.011)	(4.831)	(5.108)
Preround group output	0.501***	0.516***	0.600***	0.519*	0.352***	0.276**
	(0.0778)	(0.0913)	(0.187)	(0.277)	(0.129)	(0.129)
Constant	27.01***	47.09***	19.53**	43.27***	28.64***	44.12***
	(3.016)	(4.720)	(8.559)	(14.09)	(6.702)	(6.570)
Observations	256	256	248	248	256	256
Number of groups	32	32	31	31	32	32
R^2/Wald χ^2	653.88	179.43	108.59	82.27	37.53	57.82

Notes. Robust standard errors are in parentheses (and in (3) and (6), clustered on *group_id*). A random effects regression on periodic group output is shown.

References

Abeler J, Altmann S, Kube S, Wibral M (2010) Gift exchange and workers' fairness concerns—When equality is unfair. *J. Eur. Econom. Assoc.* 8(6):1299–1324.

Arvey RD, Murphy KR (1998) Performance evaluation in work settings. *Annual Rev. Psych.* 49:141–168.

Bellemare C, Shearer B (2009) Gift giving and worker productivity: Evidence from a firm-level experiment. *Games Econom. Behav.* 67(1):233–244.

Bernardin HJ, Cooke DK, Villanova P (2000) Conscientiousness and agreeableness as predictors of rating leniency. *J. Appl. Psych.* 85(2):232–236.

Blanco M, Engelmann D, Normann H-T (2010) A within-subject analysis of other-regarding preferences. *Games Econom. Behav.* 72(2):321–338.

Bol JC (2011) The determinants and performance effects of managers' performance evaluation biases. *Accounting Rev.* 86(5):1549–1575.

Bolton GE, Ockenfels A (2000) ERC—A theory of equity, reciprocity and competition. *Amer. Econom. Rev.* 90(1):166–193.

Bretz RD Jr, Milkovich GT, Read W (1992) The current state of performance appraisal research and practice: Concerns, directions, and implications. *J. Management* 18(2):321–352.

Charness G (2004) Attribution and reciprocity in an experimental labor market. *J. Labor Econom.* 22(3):665–688.

Charness G, Rabin M (2002) Understanding social preferences with simple tests. *Quart. J. Econom.* 117(3):817–869.

Clark AE, Masclet D, Villeval MC (2010) Effort and comparison income: Experimental and survey evidence. *Indust. Labor Relations Rev.* 63(3):407–426.

Cohn A, Fehr E, Goette L (2009) Fairness and effort: Evidence from a field experiment. Working paper, University of Zurich, Zurich.

Dannenberg 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, Zentrum für Europäische Wirtschaftsforchung, Mannheim, Germany.

Engellandt A, Riphahn RT (2011) Evidence on incentive effects of subjective performance evaluations. *Indust. Labor Relations Rev.* 64(2):241–257.

Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. *Quart. J. Econom.* 114(3):817–868.

Fehr E, Gächter S, Kirchsteiger G (1997) Reciprocity as a contract enforcement device—Experimental evidence. *Econometrica* 64(4):833–860.

Fehr E, Kirchsteiger G, Riedl A (1993) Does fairness prevent market clearing? An experimental investigation. *Quart. J. Econom.* 108(2):437–460.

Fischbacher U (2007) *z*-tree: Zurich toolbox for ready-made economic experiments. *Experimental Econom.* 10(2):171–178.

Gibbs M, Merchant KA, van der Stede WA, Vargus ME (2003) Determinants and effects of subjectivity in incentives. *Accounting Rev.* 79(2):409–436.

Gneezy U, List JA (2006) Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica* 74(5):1365–1384.

Greiner B (2004) The online recruitment system ORSEE—A guide for the organization of experiments in economics. Technical Report 2003-10, Max Planck Institute of Economics, Jena, Germany.

Hannan RL, Kagel JH, Moser DV (2002) Partial gift exchange in an experimental labor market: Impact of subject population differences, productivity differences, and effort requests on behavior. *J. Labor Econom.* 20(4):923–951.

Harbring C, Irlenbusch B (2011) Sabotage in tournaments: Evidence from a laboratory experiment. *Management Sci.* 57(4):611–627.

Hennig-Schmidt H, Rockenbach B, Sadrieh A (2010) In search of workers' real effort reciprocity—A field and a laboratory experiment. *J. Eur. Econom. Assoc.* 8(4):817–837.

Holland K (2006) Performance reviews: Many need improvement. *New York Times* (September 10), http://www.nytimes.com/2006/09/10/business/yourmoney/10mgmt.html.

Jawahar J, Williams CR (1997) Where all the children are above average: A meta analysis of the performance appraisal purpose affect. *Personnel Psych.* 50(4):905–925.

Kampkötter P, Sliwka D (2011) Differentiation and performance—An empiricial investigation on the incentive effects of bonus plans. Mimeo, University of Cologne, Cologne, Germany.

Kane JS, Bernardin HJ, Villanova P, Peyrefitte J (1995) Stability of rater leniency: Three studies. *Acad. Management J.* 38(4):1036–1051.

Kube S, Maréchal MA, Puppe C (2012) Do wage cuts damage work morale? Evidence from a natural field experiment. *J. Eur. Econom. Assoc.* Forthcoming.

Landy FJ, Farr JL (1980) Performance rating. *Psych. Bull.* 87(1):72–107.

Lazear EP (1989) Pay equality and industrial politics. *J. Political Econom.* 97(3):561–80.

Lazear EP, Rosen S (1981) Rank-order tournaments as optimum labor contracts. *J. Political Econom.* 89(5):841–864.

Moers F (2005) Discretion and bias in performance evaluation: the impact of diversity and subjectivity. *Accounting, Organ. Soc.* 30(1):67–80.

Murphy KJ (1992) Performance measurement and appraisal: Motivating managers to identify and reward performance. Burns WJ Jr, ed. *Performance Measurement, Evaluation, and Incentives* (Harvard Business School Press, Boston), 37–62.

Murphy KR, Cleveland JN (1995) *Understanding Performance Appraisal* (Sage, Thousand Oaks, CA).

Ockenfels A, Sliwka D, Werner P (2010) Bonus payments and reference point violations. IZA Discussion Paper 4795, Institute for the Study of Labor, Bonn, Germany.

Orrison A, Schotter A, Weigelt K (2004) Multiperson tournaments: An experimental examination. *Management Sci.* 50(2):268–279.

Prendergast C, Topel R (1996) Favoritism in organizations. *J. Political Econom.* 104(5):958–978.

Prendergast CJ (1999) The provision of incentives in firms. *J. Econom. Literature* 37(1):7–63.

Prendergast CJ, Topel RH (1993) Discretion and bias in performance evaluation. *Eur. Econom. Rev.* 37(2):355–365.

Rynes SL, Gerhart B, Parks L (2005) Personnel psychology: Performance evaluation and pay for performance. *Annual Rev. Psych.* 56:571–600.

Schleicher DJ, Bull RA, Green SG (2009) Rater reactions to forced distribution rating systems. *J. Management* 35(4):899–927.

Schotter A, Weigelt K (1992) Asymmetric tournaments, equal opportunity laws, and affirmative action: Some experimental results. *Quart. J. Econom.* 107(2):511–539.

Scullen SE, Bergey PK, Aiman-Smith L (2005) Forced distribution rating systems and the improvement of workforce potential: A baseline simulation. *Personnel Psych.* 58(1):1–32.