ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Berger, Johannes; Harbring, Christine; Sliwka, Dirk

Working Paper

Performance appraisals and the impact of forced distribution: An experimental investigation

IZA Discussion Papers, No. 5020

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Berger, Johannes; Harbring, Christine; Sliwka, Dirk (2010) : Performance appraisals and the impact of forced distribution: An experimental investigation, IZA Discussion Papers, No. 5020, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/36830

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

DISCUSSION PAPER SERIES

IZA DP No. 5020

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

Johannes Berger
Christine Harbring
Dirk Sliwka

June 2010

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

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

Johannes Berger
University of Cologne

Christine Harbring
Karlsruhe Institute of Technology and IZA

Dirk Sliwka
University of Cologne and IZA

Discussion Paper No. 5020
June 2010

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 5020
June 2010

ABSTRACT

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

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 were not restricted in their rating behavior to a forced distribution system in which they had to assign differentiated grades. We find that productivity was significantly higher under a forced distribution by about 8%. But also in the absence of forced distribution, deliberate differentiation positively affected output in subsequent work periods.

JEL Classification: C91, D83, J33, M52

Keywords: performance measurement, forced distribution, motivation, experiment

Corresponding author:

Christine Harbring
Karlsruhe Institute of Technology
Waldhornstr. 27
76131 Karlsruhe
Germany
E-mail: christine.harbring@kit.edu

1 Introduction

In most jobs an employee's true efforts are at best imprecisely captured by objective key figures. Hence, organizations frequently use subjective appraisals to evaluate substantial parts of an employee's job performance. While this may strengthen the setting of incentives as more facets of job performance are evaluated, the opposite may be true when supervisors bias the 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] As rating scales nearly always have an upper boundary, rater leniency often directly implies rating compression. While the existence of these biases has been confirmed in previous studies, there is surprisingly little evidence on the performance consequences of biased performance appraisals when they are tied to compensation. Rynes et al. (2005), for instance, stress 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*" (p. 572).

A simple economic logic suggests that both of the above mentioned biases should lead to weaker incentives. As 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. But on

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

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

the other hand, 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 distribution" systems under which supervisors have to follow a predetermined distribution of ratings. At General Electric for example, the former CEO 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. According to estimates, a quarter of the Fortune 500 companies (e.g. Cisco, Intel, Hewlett Packard, Microsoft etc.) link parts of individual benefits to a relative performance evaluation (Boyle (2001)). However, the use of these systems is often very controversially discussed and in some firms even led to lawsuits as employees claimed to have been treated unfairly.[4]

A key reason for the lack of field evidence on the consequences of a forced distribution is that even when a firm changes its system of performance appraisals there is typically no control group within the same firm with an unaltered scheme which in turn makes it hard to identify the causal effect of the modification. Moreover, to measure the performance consequences an objective measure of individual performance is necessary. But such objective measures are typically not available when subjective assessments are used.[5]

[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), Fehr et al. (1997), Hannan et al. (2002) or Charness (2004). Evidence from field experiments is somewhat less pronounced. Recent studies find mostly moderate support for positive reciprocity. See for instance Gneezy and List (2006), Cohn et al. (2009), Kube et al. (2010), Bellemare and Shearer (2009) and Hennig-Schmidt et al. (2010).

[4]See for instance "Performance Reviews: Many Need Improvement" in the New York Times (September 10, 2006).

[5]Typical examples of departments in which objective measures of performance are available are sales functions in which revenues of individual sales agents can be measured. But in these departments subjective assessments and in particular forced distributions are hardly ever used because the objective performance measures already lead to differentiated ratings.

Hence, 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 while the outcome of their work directly determines the supervisor's payoffs. At the end of each round the supervisor learns the work outcome of each individual employee and is then asked to individually rate their performance on a five point scale. The employees 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 after some rating experience with or without such a system.

Our key result is that worker productivity in our experiment is by about 8% 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. If, for instance, an employee receives the best potential rating but does not have the highest work outcome in the group, his subsequent performance decreases. 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 as compared to a setting in which it has never been used. But, on the other hand, the performance effect of a forced distribution is strongly reduced when the participants have experienced the more "liberal" baseline setting before and, hence, have different reference standards and expectations.

While 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 (2008), Bol (2009), Kampkötter and Sliwka (2010) and Ahn et al. (2008) give 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) have experimentally investigated rater's reaction to forced distribution and find 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) conduct a simulation study and show that forced distribution can increase performance in the short run as low performers are driven out of the firm. This effect, however, becomes smaller over time. Neither study examines the incentive effects of forced distributions.

The paper proceeds as follows. In the next section the experimental design and procedure are described. The experimental results are summarized in section 3. We first provide evidence on the performance difference between our baseline treatment and the forced distribution condition. Then, we take a closer look at rating decisions within the baseline treatment and their relation to workers' performance as well as the connection between the supervisor's social preferences and rating behavior. Finally, we investigate the effect of past experience in a different rating setting on both the supervisor's and the workers' behavior. We discuss and conclude our results in the last section.

2 Experimental Design

We conduct a real effort laboratory study. The majority of the subjects has to work on a tedious task in the role of a "worker". Their individual performance is evaluated by other subjects who are assigned the role of a "supervisor". The experiment consists of several parts which are described in the following.

Ability Test

In an initial pre-round all subjects have to work on the real effort task which is also used in the main part of the experiment, i.e. all have to repeatedly count the number "7" in blocks of randomly generated numbers. This pre-round is conducted to collect a measure for each subject's ability for the task and also to familiarize participants with the task (also those who are in the role of the supervisor). To make sure everybody has correctly understood the task, an "exercise block" is presented on the computer screen prior to the pre-round. Only after all subjects have correctly solved this block, the pre-round which lasts for 2.5 minutes is started. During the pre-round subjects' performance is measured by the number of 'points' they collect which is converted into Euro after the experiment. For each correct answer a subject receives two points, for each wrong answer it loses 0.5 points. At the end of the round, a piece-rate of 10 cents per point is paid to each participant's account. During the task subjects are also offered the opportunity to use a "time-out" button which locks the screen for 20 seconds during which subjects cannot work on any blocks. Each time the time-out button is pushed the subject receives 8 cents. This time-out button is implemented to simulate potential opportunity costs of working. At the end of the pre-round each participant is informed about the total number of points achieved as well as the number of correct and false answers and the resulting payoff.[6]

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 is started, participants have to answer several test questions on the screen to make sure that they have fully understood the procedure and the calculation of the payoffs.[7] This first

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

[7] Participants had to calculate the payoffs for a worker and a supervisor for an output as well as a rating they themselves could freely choose.

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 pre-round. They can again make use of a time-out button blocking the screen for 20 seconds for which they receive 25 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 all 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.

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

Table 1: Ratings and Bonus Payments

Each rating is associated with a bonus payment for the worker (see table 1) ranging from 10 € for the highest rating "1" to 0 € for the worst rating of "5". 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 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. One round is randomly determined in each part of the experiment which is payoff-relevant (for details see "Procedures").

Matching of Groups

To create a situation in which performance ratings are not straightforwardly due to ability differences, we match participants into homogeneous groups. The matching procedure is based on the performance in the pre-round, i.e. all 32 subjects are individually ranked in each session based on their total number of points achieved in the pre-round. 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 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 to elicit subjects' social preferences. After these games all participants have to fill out a questionnaire.

Treatments

We analyze two different settings: In the Baseline setting (*Base*) supervisors are not restricted in their rating behavior. In the Forced Distribution setting (*Fds*), however, supervisors have to give one worker a rating of "1" or "2", one worker a rating of "3" and another worker a "4" or "5". This restriction is explained to all participants in the treatment.

To also 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 8 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 8 rounds (first part) which are followed by 8 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*). Therefore, we conduct four treatments in total (see table 2).

Treatment	Round 1-8	Round 9-16
BaseBase	Base	Base
FdsFds	Fds	Fds
BaseFds	Base	Fds
FdsBase	Fds	Base

Table 2: Overview Treatments

Procedures

After participants have arrived in the laboratory, they are seated in separated cabins where they receive the instructions for the pre-round of the experiment. Participants are advised that they are not allowed to communicate. In case of any question they have to raise their hand such that one of the experimenters will come and help. The experiment starts after all participants have read the instructions and all questions have been answered. After the pre-round, instructions for the first part of the experiment are distributed. Instructions for the second part only follow after the first part has been completed.[8]

The instructions inform participants that only one of the eight rounds of each part of the experiment will be payoff-relevant for all participants. At the end of each session a randomly selected subject is asked to twice draw one of 8 cards to determine which rounds will be paid out. The final payoff for each subject consists of the money earned during the experiment and a show-up fee of 4 € . The money is anonymously paid out in cash at the end of each session.

In total the experiment consists of 8 sessions with two sessions for each treatment condition. Thus, we have 64 subjects (16 independent groups) in

[8]In *BaseBase* and *FdsFds* the subjects are told after the first part that the rules for the second part of the experiment are the same as for the first part.

each treatment with a total of 256 participants. It is ensured that no one has been involved in an experiment with the same real effort task before. No subject participates in more than one session. On average a session lasts for 2.5 hours and the average payoff amounts to 27 €. The experiment is conducted at the Cologne Laboratory for Economic Research. All sessions are computerized using the experimental software z-Tree (Fischbacher (2007)) and subjects are 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 by comparing the treatments *BaseBase* and *FdsFds*. We then analyze the driving forces behind the observed differences in more detail. Finally, we provide an overview of spillover effects observed when varying the sequence of both settings in *BaseFds* and *FdsBase*.

3.1 Performance Effects of Forced Distribution

Figure 1 depicts the distribution of ratings in *BaseBase* and *FdsFds*. Evidently, supervisors tend to assign very good ratings, i.e. a "1" or "2" in the majority of cases in *BaseBase* (83%). Note that this pattern closely resembles the typical "leniency bias" often observed in organizational practice. Bretz et al. (1992), for instance, describe this as follows: "*Performance appraisal systems typically have five levels to differentiate employee performance. However, even though most organizations report systems with five levels, generally only three levels are used. Both the desired and the actual distributions tend to be top heavy, with the top "Buckets" relatively full and the bottom buckets relatively empty... It is common for 60-70% of an organization's workforce to be rated in the top two performance levels. ... Skewed performance distributions not only exist, but are common*". As is the case

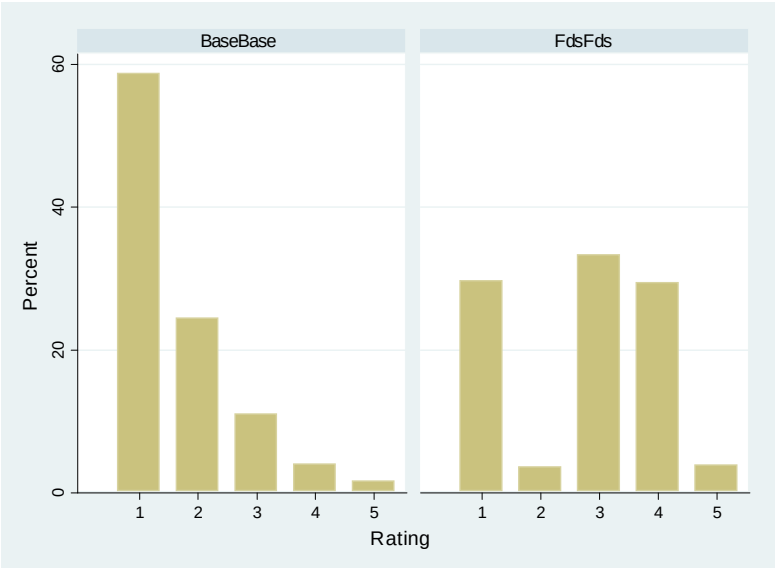

Figure 1: Distribution of Ratings in *BaseBase* and *FdsFds*

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 obviously 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 our experiment still follow the lenient choices and strongly prefer the "1" over the "2" and the "4" over the "5" as shown in the right panel of figure 1.

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

numerous gift-exchange experiments have now provided evidence for the fair wage effort hypothesis by Akerlof and Yellen (1990) showing that higher wage payments indeed may lead to higher efforts.

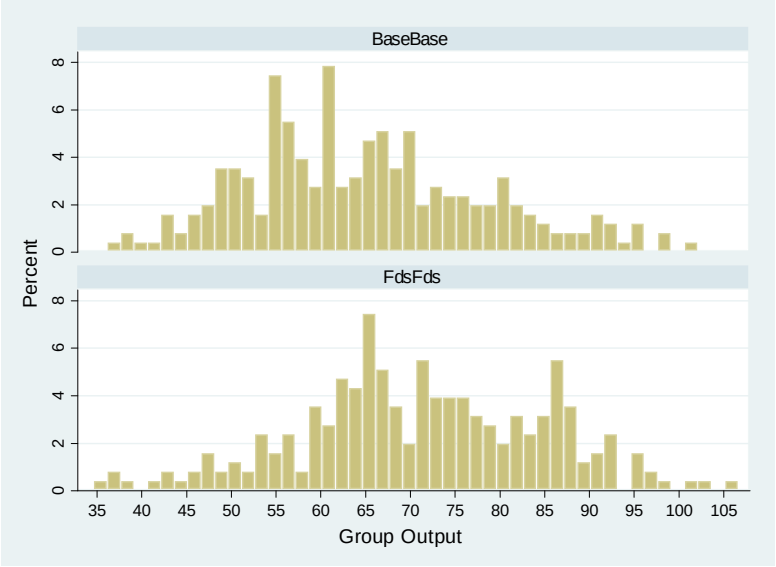

Figure 2: Distribution of Output in *BaseBase* and *FdsFds*

Figure 2 displays the distribution of outputs in both treatments. The figure indicates that performance indeed seems to be higher under the forced distribution. Average group output amounts to 65.02 in *BaseBase* and to 71.38 in *FdsFds*. Our matching procedure is designed to generate groups of similar ability in both treatments. Thus, to give an accurate test without any distributional assumptions on whether the apparent performance difference is statistically significant we have to compare the groups across treatments according to their rank in the ability test. Hence, we pair the highest ranking group according to the ability test in *BaseBase* with the highest ranking group in *FdsFds* and then the second highest groups from both treatments and so on. We then counted in how many of these pairs the average group performance across all rounds is higher in the *FdsFds* treatment. As this

is the case in 12 of the 16 pairs, the difference is significant ($p = 0.038$, one-sided Binomial test).

We further investigate this performance difference by running random-effects regressions with the group output (sum of the points of all 3 workers) and alternatively the logarithm of group output in a period as the dependent variable. Due to the matching procedure we have to control for innate group ability measured by the group output from the initial pre-round. The results are reported in table 3. As model (1) indicates group output is about 5 points higher under forced distribution. Model (5) shows that this translates into an average productivity difference of 8%. Models (2) and (3) show that this holds for both parts of the experiment. Model (4) adds an interaction term between the *Fds* dummy variable and group ability. The significant negative coefficient indicates that productivity enhancement is particularly high for groups with a low output in the pre-round.

Dependent Variable:	Group Output				Log Group Output			
	BaseBase vs. FdsFds				BaseBase vs. FdsFds			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Periods	Periods	Periods	Periods	Periods	Periods	Periods	Periods
	1-16	1-8	9-16	1-16	1-16	1-8	9-16	1-16
Fds	5.257**	4.728**	5.786**	18.98***	0.0781**	0.0723**	0.0839**	0.296***
	(2.208)	(2.155)	(2.456)	(6.658)	(0.033)	(0.034)	(0.035)	(0.099)
Group Ability	0.508***	0.501***	0.516***	0.696***	0.0075***	0.0078***	0.0071***	0.0105***
	(0.082)	(0.078)	(0.091)	(0.092)	(0.001)	(0.001)	(0.001)	(0.001)
Fds × Group Ability				-0.294**				-0.0047***
				(0.123)				(0.002)
Constant	26.40***	27.01***	47.09***	17.77***	3.561***	3.548***	3.907***	3.424***
	(3.362)	(3.016)	(4.720)	(4.740)	(0.051)	(0.048)	(0.071)	(0.067)
Observations	512	256	256	512	512	256	256	512
Number of Groups	32	32	32	32	32	32	32	32
Wald Chi^2	1640.10	653.88	179.43	1487.50	1458.06	597.49	192.19	1613.19

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

Random effects regression (period dummies included)

Table 3: The Impact of Forced Distribution on Productivity

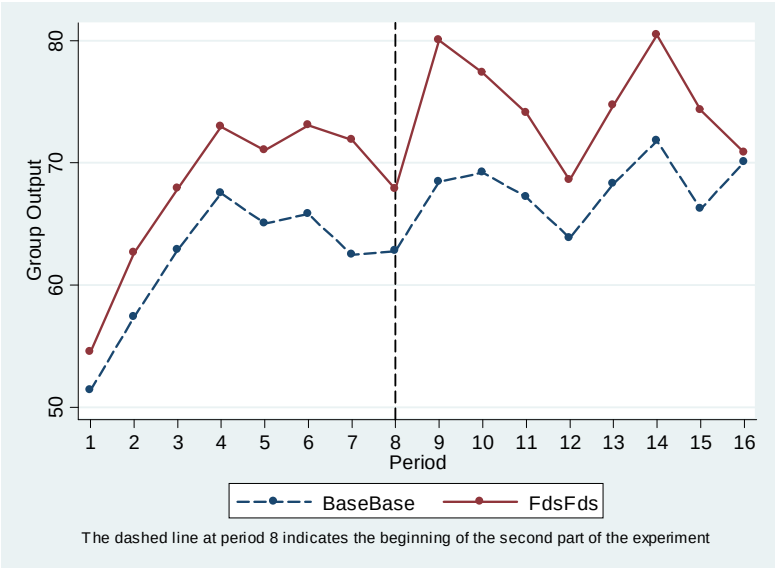

Figure 3: Output in *BaseBase* and *FdsFds* over Time

Taking a closer look at the evolution of work performance over time in Figure 3, we see that performance differences become larger in the second part of the experiment. Hence, participants seem to have learned over the course of the experiment that performance incentives are stronger under the forced distribution.[9] The increase in productivity under *Fds* is also reflected by a considerable decrease in the use of the time-out button in the second part of the experiment. In only 8 out of 128 cases (6%) time-outs are observed during the second part of the experiment under a forced distribution while this is true for 48 out of 128 cases (38%) in the baseline setting.[10]

Investigating the treatment differences with alternate productivity measures, such as the number of blocks finished per group and the number of

[9]It is interesting to note that the qualitative shape of both graphs over time is quite similar reflecting parallel effects of learning and fatigue.

[10]Note that even after excluding all periods in which subjects take timeouts, performance is still significantly higher under *Fds*. This suggests that subjects not only stop working more frequently but that they also exert less effort while working on the task in the baseline compared to the forced distribution.

correct and false answers (see table A1 in the appendix) we find that under forced distribution subjects count and solve more blocks correctly while making only slightly and insignificantly more mistakes.

3.2 Differentiation and Productivity

But why do people work harder under the forced distribution? A key conjecture is that under the forced distribution supervisors differentiate more according to individual performance which strengthens the incentives to exert effort. 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 or 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 as they are not forced to differentiate. On the other hand, variations in absolute performance may affect the grades in both treatments. To investigate this we run random effects regressions with the bonus received in a period as dependent and the absolute output and relative rank as independent variables.[11] To test for treatment differences we include interaction terms with a dummy variable for the forced distribution treatment.

The results are reported in table 4. Note that the relative rank matters in both treatments but does so to a much larger extent under forced distribution. Interestingly, while within-rank variation in output is rewarded in both treatments, these rewards are stronger in the baseline treatment, i.e., for a given rank output and bonus are more strongly (positively) correlated in the baseline treatment. But, apparently, competing for ranks generates stronger incentives in the forced distribution treatment.[12]

[11] The last rank 3 is the reference group.

[12] The competition for ranks indeed induces a 'tournament' among the agents. As the

Dependent Variable:	Individual Bonus	
	BaseBase	
	(1)	(2)
	Periods	Periods
	1-8	9-16
Output	0.281***	0.275***
	(0.026)	(0.029)
Output × Fds	-0.216***	-0.221***
	(0.022)	(0.020)
Rank 2	0.802***	0.698***
	(0.252)	(0.230)
Rank 1	1.104**	0.840**
	(0.436)	(0.327)
Rank 2 × Fds	1.242***	1.734***
	(0.353)	(0.259)
Rank 1 × Fds	4.712***	5.638***
	(0.626)	(0.429)
Individual Ability	-0.0650***	-0.0585**
	(0.022)	(0.024)
Constant	3.173***	2.278***
	(0.536)	(0.622)
Observations	768	768
Number of Subjects	96	96
Wald Chi^2	743.26	2970.26

Robust standard errors in parentheses (clustered on group_id)
*** p<0.01, ** p<0.05, * p<0.1
Random effects regression (period dummies included)

Table 4: The Impact of Rank and Output on Bonus Payments

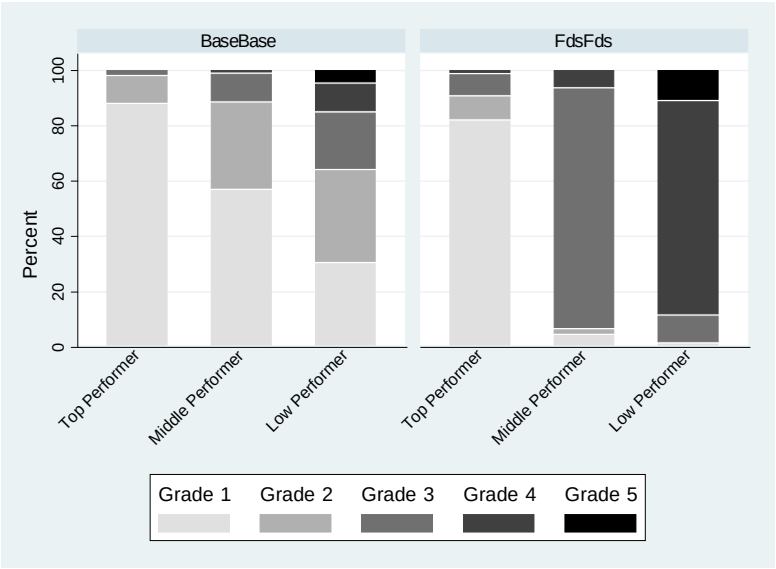

Figure 4: Distribution of Ratings according to Relative Performance in the Group

Figure 4 shows the distribution of grades for the top, middle, and low performers.[13] In the forced distribution treatment 90% of the participants with the highest rank receive a 1 or a 2 and 89% with the lowest rank a 4 or a 5. In contrast, in the baseline treatment about 64% of the worst performers still receive a 1 or a 2.

Hence, the gains from improving the rank are much weaker in the baseline treatment. Thus, it seems important to analyze whether these weaker benefits from improving one's relative position can indeed explain lower outputs in the baseline setting. We, therefore, investigate the impact of particular grades on individuals' subsequent performance in the baseline treatment.

literature on tournaments starting with Lazear and Rosen (1981) has shown, they can indeed be powerful incentive instruments. For experimental evidence on tournaments see for example Schotter and Weigelt (1992), Orrison et al. (2004) or Harbring and Irlenbusch (2009).

[13] We define top, middle and low performers according to the relative performance rank in the group in a given round.

Table 5 reports results from a random effects regression in the baseline treatment with individual output in $t+1$ as the dependent variable and dummy variables for the grade assigned in period t as independent variables. The reference category corresponds to receiving the top grade "1". Analyzing the reaction of all workers in model (1) we find that obtaining the medium grade "3" instead of a "1" has a significant positive effect on output in the subsequent round. Model (2) and (3) disentangle the overall effect according to the participants' rank. Model (2) only includes the observations of the top performers in each period,[14] model (3) only the observations of the other subjects, i.e., middle and low performers. Interestingly, top performers do not adjust their effort after receiving a lower grade instead of the top grade. However, middle and low performers substantially increase their outputs when receiving a "2" or a "3" compared to receiving the top grade "1".[15] Thus, those who know that they attained the highest output are not motivated by getting lower ratings. In contrast, those who are not the best performers and yet receive the top grade reduce their efforts which supports the view that lenient and undifferentiated ratings indeed undermine performance incentives.[16]

These results suggest that supervisors will induce higher performance in subsequent rounds by using a larger span of grades. To test this we run random effects regressions in *BaseBase* using the group output in period $t+1$ as the dependent variable and dummy variables for each span of grades, i.e. the difference between the worst and the best rating assigned by the supervisor, in round t as key independent variables. The results are reported in table 6. No differentiation, i.e. cases in which each worker receives the same

[14] Note that the model identifies the effects of grading by comparing situations in which a given person obtained the highest rank but received different grades.

[15] Note that only a few top performers received a "3". Moreover, even to the middle and low performers "4" and "5" were rarely assigned.

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

Dependent Variable:	Output_{t+1}		
	BaseBase		
	(1)	(2)	(3)
	All Workers	Top	Middle/Low
	Periods	Periods	Periods
	1-16	1-16	1-16
$\text{Grade}=2_t$	0.633	-0.0612	1.008**
	(0.492)	(1.067)	(0.413)
$\text{Grade}=3_t$	2.317***	1.769	2.182***
	(0.451)	(1.722)	(0.422)
$\text{Grade}=4_t$	1.149		0.801
	(1.285)		(1.450)
$\text{Grade}=5_t$	3.631		2.029
	(2.213)		(2.601)
Output_t	0.639***	0.646***	0.474***
	(0.061)	(0.060)	(0.101)
Individual Ability	0.296***	0.460***	0.317***
	(0.034)	(0.108)	(0.065)
Constant	3.013***	1.450	4.811***
	(0.954)	(1.659)	(1.036)
Observations	720	260	460
Number of Subjects	48	42	47

Robust standard errors in parentheses (clustered on group_id)

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

Random effects regression (period dummies included)

Table 5: The Impact of Ratings on Individual Performance

rating, serves as our reference category.[17] The results suggest that extending the range of applied ratings from 0 to 2, for instance, increases subsequent productivity on average by almost 3 points (4%) in the first part and more than 6 points (9%) in the second part.[18]

Additional evidence for positive effects of deliberate differentiation can be derived from our post-experimental questionnaire. As already mentioned above we ask subjects in the role of the supervisors about their rating behavior in both parts of the experiment. The items[19] "I assigned bad ratings to motivate the workers" and "I assigned bad ratings to sanction the workers" are both positively correlated with higher a group output in the second part of the experiment (significant at the 10% and 5%-level). A regression analysis shows that this is still true after controlling for group ability (see regression table A2 in the Appendix). Moreover, these self reported measures of differentiation are highly correlated with actual differentiation in the second part of the experiment (e.g. range of grades or standard deviation of grades) controlling for group output.

3.3 Social Preferences and Differentiation

As has already been revealed by psychological studies there is some evidence that the personality of the supervisor matters for the evaluation behavior.[20] 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 the way in which performance ratings are assigned. To investigate this we elicit subjects' social preferences before final payoffs

[17] In 27% of all rounds in *BaseBase* the supervisor assigned all workers the best rating "1" and in 29% of all rounds she/he assigned the same rating to all three participants.

[18] Note that an observed range of grades larger than 2 occured in only 30 out of 256 rating decisions in the baseline treatment.

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

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

Dependent Variable:	Group Output$_{t+1}$		
	BaseBase		
	(1)	(1)	(2)
	Periods	Periods	Periods
	1-16	1-8	9-16
Span of Grades=1$_t$	3.408***	4.501***	2.737
	(1.030)	(1.364)	(1.719)
Span of Grades=2$_t$	4.032***	2.855**	6.366***
	(1.140)	(1.394)	(1.405)
Span of Grades=3$_t$	1.038	-0.633	6.107
	(2.920)	(4.646)	(3.779)
Span of Grades=4$_t$	1.959	-1.508	6.065***
	(3.072)	(4.563)	(2.034)
Group Output$_t$	0.388***	0.432***	0.331***
	(0.0641)	(0.0776)	(0.0851)
SD of Output$_t$	0.0127	-0.125	-0.0662
	(0.250)	(0.269)	(0.242)
Group Ability	0.443***	0.395***	0.511***
	(0.094)	(0.107)	(0.119)
Constant	13.86***	13.97***	20.22***
	(2.415)	(2.933)	(3.607)
Observations	240	112	112
Number of Groups	16	16	16
Wald Chi^2	.	1195.671	4493.683

Robust standard errors in parentheses

*** p < 0.01, ** p < 0.05, * p < 0.1

Random effects regression (period dummies included)

Table 6: The Impact of Deliberate Differentiation on Subsequent Output

are communicated in our experiment. We apply an incentivized experimental procedure introduced by Blanco et al. (2007) and modified by Dannenberg et al. (2007). This simple two-step procedure, for instance, allows the experimenter to measure subjects' preferences for equity in the well-known Fehr and Schmidt (1999) utility model, according to which the utility of a person i may not only depend on her own payoff x_i but also on the difference to the payoffs of other individuals. In a two-person case it is given by

$$u(x_i, x_j) = x_i - \alpha \max\{x_j - x_i, 0\} - \beta \max\{x_i - x_j, 0\}.$$

Hence, α measures the degree to which an individual suffers from disadvantageous inequity ("envy") and β the aversion to advantageous inequity ("compassion").[21] But as has been stressed for instance by Charness and Rabin (2002) many individuals are also motivated by efficiency concerns, i.e. they may strive for maximizing the total surplus of all individuals to some extent. As laid out by Blanco et al. (2007) (footnote 33 on p.33) the Fehr Schmidt utility function also captures surplus concerns: An extended utility function $x_i - \alpha' \max\{x_j - x_i, 0\} - \beta' \max\{x_i - x_j, 0\} + \gamma(x_i + x_j)$ which allows that the agent also cares for the total surplus can be transformed into a standard Fehr-Schmidt utility function in which higher surplus concerns (γ) simply lead to a weaker disutility from disadvantageous inequality and a stronger disutility from advantageous inequality.[22]

We now expect that more "compassionate" supervisors (i.e. those with higher values of β) who care more for the well-being of the agents should assign more lenient ratings. As the rating scale is bounded, this should also lead to a weaker performanced-based differentiation. On the other hand, supervisors with higher αs are more "envious" (or less interested in the surplus

[21] See table A3 for the specific procedure used in this experiment.

[22] To be more specific, the function is equivalent to an affine transformation of $x_S - \frac{\alpha'-\gamma}{1+2\gamma} \max\{x_j - x_i, 0\} - \frac{\beta'+\gamma}{1+2\gamma} \max\{x_i - x_j, 0\}$. Hence, the experimental procedure directly yields estimates for $\alpha = \frac{\alpha'-\gamma}{1+2\gamma}$ and $\beta = \frac{\beta'+\gamma}{1+2\gamma}$.

of the agents) and should be less lenient. Moreover, a supervisor with a high α may dislike to pay an agent a bonus which is higher than her own earnings from this agent's efforts. In turn we might expect that high α individuals choose ratings which are to a stronger extent performance-contingent.

Table 7 depicts a regression analysis of different measures of leniency and rating differentiation (average grade in model (1), the range of grades in (2), the standard deviation of grades in (3), the coefficient of variation of grades in (4) and the probability that all receive the best grade "1" in (5) on our proxies for inequity aversion). In all of our specifications we control for the sum and standard deviation of outputs as observed by the supervisor. Furthermore, as we pool all our four treatments, we include treatment dummies and interact the inequity parameters with our treatment dummy variable *Fds*.

The results are well in line with the hypotheses and surprisingly robust. Supervisors with higher betas and lower alphas indeed give better grades, differentiate less, and assign the best grade to all their agents with a higher probability. The interaction terms show that the effects of social preferences disappear under the forced distribution. Hence, a forced distribution system indeed avoids biases due to differing degrees of the supervisors' social preferences.

Dependent Variable:	Mean Grade	Range of Grades	SD of Grades	CV Grades	All 1
	All Treatments pooled				
	(1)	(2)	(3)	(4)	(5)
Alpha	0.162**	0.123**	0.229**	0.0605***	-0.109***
	(0.074)	(0.058)	(0.111)	(0.022)	(0.040)
Beta	-0.309*	-0.273**	-0.507**	-0.122***	0.209**
	(0.163)	(0.115)	(0.226)	(0.044)	(0.086)
Fds	0.933***	0.668***	1.441***	0.121***	-0.138***
	(0.072)	(0.085)	(0.162)	(0.034)	(0.049)
Fds × Alpha	-0.0634	-0.143**	-0.264**	-0.0761***	0.0769*
	(0.066)	(0.067)	(0.127)	(0.030)	(0.043)
Fds × Beta	0.126	0.255**	0.469*	0.127**	-0.154*
	(0.117)	(0.127)	(0.248)	(0.053)	(0.081)
Group Output	-0.0156***	-0.0080***	-0.0151***	-0.0022***	0.0051***
	(0.002)	(0.002)	(0.004)	(0.001)	(0.001)
SD of Output	0.0198***	0.0578***	0.110***	0.0218***	-0.0125***
	(0.005)	(0.006)	(0.011)	(0.002)	(0.003)
Constant	2.651***	0.976***	1.800***	0.427***	-0.113
	(0.204)	(0.125)	(0.241)	(0.049)	(0.082)
Observations	976	976	976	976	976
Number of Groups	61	61	61	61	61
Wald Chi^2	2186.75	909.43	1040.10	434.96	132.55

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

Random effects regression (period and treatment dummies included)

Table 7: Inequity Aversion and Rating Behavior in the Experiment

We also test the relation between inequity aversion and the tendency to compress ratings non-parametrically in our baseline treatment. After dividing supervisors at the median of the beta distribution into two groups, we use the Mann-Whitney U-test to explore differences in rating behavior. Even though we do not control for differences in output, we have marginally significant results that the rating behavior is different across these two groups.[23]

3.4 Introducing or Abolishing a Forced Distribution?

In this last section we take a closer look at effects of a within treatment variation in the rules of performance evaluation. As 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. Because we have to take learning effects into account we compare the performance with the treatment in which the baseline setting is played in both parts (i.e. the second part of the *BaseFds* treatment with the second part of *BaseBase*).

Given the results of the between treatment comparison described above, we should expect an increase in performance after the forced distribution is introduced. However, 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 8. However, a surprising pattern emerges when we compare the effects per period as shown in column (2). While performance increases by about 5 points in period 9, the first period after the forced distribution has been introduced, and stays at this level in period 10, it drops to a level roughly 2-3 points below the baseline level in the last 6 periods. Hence, participants are apparently initially motivated

[23] We apply a two-tailed Mann-Whitney U-test to the different measures of rating differentiation to compare groups of supervisors whose betas are above the median of all supervisors' betas in the treatment to those whose betas are below the median: Mean grade: p=0.138, range of grades: p=0.0899, standard deviation of grades: p=0.0901, coeffient of variation: p=0.0807, frequency of the mean grade=1: p=0.0699.

to work harder under the forced distribution as 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. But in contrast to a setting in which a forced distribution is present from the outset, the participants now have a different reference standard as they have already experienced more favorable ratings, which may cause their reduced motivation. This is 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 potential explanation for this observation 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 *BaseFds* treatment to the treatment in which the forced distribution has been used throughout the experiment. But as column (1) of table 9 shows, the forced distribution system in the second part performs worse after the baseline setting as compared to the situation in which agents work under a forced distribution right from the beginning. Hence, it is indeed the experience of the baseline setting with higher grades and bonuses which leads to a demotivational effect of the forced distribution. The negative perception of this relative loss of payments apparently seems to counteract the positive forces of increased differentiation.

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) indicates that the performance difference to the second part of *BaseBase* amounts to roughly 7% on average. Analogously to the above reasoning workers in *FdsBase* seem to be particularly motivated in the second part as 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 8 rounds *(BaseBase)* the workers in *FdsBase* could, thus, feel more inclined to reciprocate this relative increase

Dependent Variable:	Group Output$_t$	
	BaseFds vs. BaseBase	
	(1)	(2)
	Periods 9-16	Periods 9-16
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)
Group Ability	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 Chi^2	148.70	325.12

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Random effects regression (period dummies included)

Table 8: Effects of the Introduction of a Forced Distribution (BaseFds vs. BaseBase, periods 9-16)

in bonus payments. Yet, another factor driving this result is that supervisors keep up differentiation even after forced distribution has been abolished. Indeed, we find some evidence that supervisors in *FdsBase* tend to differentiate more during the second part than their counterparts in *BaseBase*. Workers ranked 2^{nd} or 3^{rd} in a group are significantly less likely to receive a "1" for a given output and more likely to receive a "4" or "5" in second part of *FdsBase* than in *BaseBase* (see table A4). Hence, the experience with a forced distribution apparently has helped to establish a norm of making performance-contingent ratings which indeed leads to a better performance.

Dependent Variable:	Group Output	
	BaseFds vs. FdsFds	BaseBase vs. FdsBase
	(1)	(2)
	Periods	Periods
	9-16	9-16
BaseFds	-5.763*	
	(2.994)	
FdsBase		4.591*
		(2.363)
Group Ability	0.514***	0.644***
	(0.087)	(0.105)
Constant	56.09***	41.04***
	(5.187)	(5.312)
Observations	256	256
Number of Groups	32	32
Wald Chi^2	318.06	57.95

Robust standard errors in parentheses, *** p < 0.01, ** p < 0.05, * p < 0.1
Random effects regression (period dummies included)

Table 9: Introducing and Abolishing Forced Distribution

Additional evidence for these arguments comes from our post-experimental questionnaire. We pose participants who experience both settings in *BaseFds* and *FdsBase* a variety of questions separately for both parts of the exper-

iment. Especially workers in *BaseFds* feel that their effort pays more off and that their well-being is more important to the supervisor during the baseline setting. They also state that the supervisor's behavior is more fair and that he is more capable of giving appropriate ratings in the absence of *Fds*.[24] The supervisors naturally also express some dissatisfaction towards the forced distribution as, for instance, they feel the rating decision to be more difficult in the second part of *BaseFds* which is well in line with the findings by Schleicher et al. (2009).

4 Conclusion

We study 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 by about 8% higher under the forced distribution in a between-subjects design. The reason for this substantial gain in performance is that many supervisors in the baseline setting seem to be too lenient in their rating decisions and, hence, performance incentives are too weak. But even within the baseline setting those supervisors who choose less lenient and more differentiated ratings attain a higher performance. The forced distribution system creates stronger performance incentives across all supervisors.

Moreover, we analyze potential effects of the supervisor's social preferences on rating behavior. We find that a supervisor's social preferences have a substantial impact on her rating behavior in the baseline setting. More "compassionate" supervisors who care more for the well-being of the agents indeed assign more lenient ratings, and more "envious" supervisors do the opposite. These differences vanish under a forced distribution.

But our results also indicate that it may be problematic to set up a forced

[24] We apply the Wilcoxon-Signed Rank test for dependent pairs for the answers of each subject related to the two settings *Base* and *Fds* in *BaseFds*. All differences reported here are significant at a level of at least 10%, two-tailed. Differences in survey answers in *FdsBase* are quite similar but not consistently significant.

distribution when employees have experienced a more "liberal" system of performance evaluations before. Most importantly, in our within-subjects design we find that the introduction of a forced distribution leads to a short-term performance increase which is followed by a rather sharp drop in performance. Apparently, while the participants initially understand that they need to work harder under a forced distribution they are soon demotivated as they cannot attain the good grades and high bonuses they have earned before. On the other hand, some experience with the forced distribution in the beginning demonstrates supervisors the benefits of differentiation as they tend to differentiate more and attain a higher performance even in the baseline setting as compared to supervisors without the experience of a forced distribution.

Our results have several interesting implications for the design of performance evaluation schemes in practice. First of all, forced distribution systems may indeed lead to performance increases as sometimes conjectured by practitioners. However, our results also show that "history matters", i.e. 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 have been shaped by their previous experience and the way in which appraisals have been assigned in the past. But these reference standards carry over to the new system and affect the social, economic and psychological mechanisms at work in the appraisal process.

References

Abeler, J., S. Altmann, S. Kube, and M. Wibral (2010). Gift exchange and workers' fairness concerns - when equality is unfair. *Journal of the European Economic Association (forthcoming)*.

Ahn, T. S., I. Hwang, and M.-I. Kim (2008). Discriminability of subjective performance measures and its impact on ratee incentives. *Working Paper*.

Akerlof, G. A. and J. L. Yellen (1990). The fair wage-effort hypothesis and unemployment. *The Quarterly Journal of Economics 105*(2), 255–83.

Arvey, R. and K. Murphy (1998). Performance evaluation in work settings. *Annual Review of Psychology 49*(1), 141–168.

Bellemare, C. and B. Shearer (2009). Gift giving and worker productivity: Evidence from a firm-level experiment. *Games and Economic Behavior 67*(1), 233 – 244. Special Section of Games and Economic Behavior Dedicated to the 8th ACM Conference on Electronic Commerce.

Bernardin, H., D. Cooke, and P. Villanova (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology 85*, 232–236.

Blanco, M., D. Engelmann, and H. Normann (2007). A within-subject analysis of other-regarding preferences. *Manuscript, Royal Holloway*.

Bol, J. C. (2009). The determinants and performance effects of supervisor bias. *Working Paper*.

Boyle, M. (2001). Performance reviews: Perilous curves ahead. *Fortune 143*, 187–188.

Bretz, R. D. J., G. T. Milkovich, and W. Read (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management 18*, 321–352.

Charness, G. (2004). Attribution and reciprocity in an experimental labor market. *Journal of LaborEconomics 22*, 665–688.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics 117*, 817–869.

Clark, A. E., D. Masclet, and M. C. Villeval (2010). Effort and comparison income: Experimental and survey evidence. *Industrial and Labor Relations Review 63*(3), 407–426.

Cohn, A., E. Fehr, and L. Goette (2009). Fairness and effort: Evidence from a field experiment. *Working Paper*.

Dannenberg, A., T. Riechmann, B. Sturm, and C. Vogt (2007). Inequity aversion and individual behavior in public good games: An experimental investigation. *Working Paper*.

Engellandt, A. and R. T. Riphahn (2008). Incentive effects of bonus payments: Evidence from an international company. *Industrial Labor Relation Review forthcoming*.

Fehr, E., S. Gächter, and G. Kirchsteiger (1997). Reciprocity as a contract enforcement device - experimental evidence. *Econometrica 64*, 833–860.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993). Does fairness prevent market clearing? an experimental investigation. *Quarterly Journal of Economics 108*, 437–460.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics 114*, 817–868.

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

Gibbs, M., K. A. Merchant, W. A. van der Stede, and M. E. Vargus (2003). Determinants and effects of subjectivity in incentives. *The Accounting Review 79*, 409–436.

Gneezy, U. and J. A. List (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.

Hannan, R. L., J. H. Kagel, and D. V. Moser (2002). Partial gift exchange in an experimental labor market: Impact of subject population differences, productivity differences, and effort requests on behavior. *Journal of Labor Economics 20*(4), 923–951.

Harbring, C. and B. Irlenbusch (2009). Sabotage in tournaments: Evidence from a laboratory experiment. *IZA Discussion Paper No. 4205*.

Hennig-Schmidt, H., B. Rockenbach, and A. Sadrieh (2010). In search of workers' real effort reciprocity - a field and a laboratory experiment. *Journal of the European Economic Association (forthcoming)*.

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

Kampkötter, P. and D. Sliwka (2010). Differentiation and performance - an empiricial investigation on the incentive effects of bonus plans. *mimeo*.

Kane, J. S., H. J. Bernardin, P. Villanova, and J. Peyrefitte (1995). Stability of rater leniency: Three studies. *Academy of Management Journal 38*(4), 1036 – 1051.

Kube, S., M. A. Maréchal, and C. Puppe (2010). Do wage cuts damage work morale? evidence from a natural field experiment. *IEW Working Paper No. 377*.

Landy, F. J. and J. L. Farr (1980). Performance rating. *Psychological Bulletin 87*, 72–107.

Lazear, E. P. and S. Rosen (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy 89*, 841–864.

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

Murphy, K. J. (1992). Performance measurement and appraisal: Motivating managers to identify and reward performance. In W. J. J. Burns (Ed.), *Performance Measurement, Evaluation, and Incentives*, Boston, MA, pp. 37–62. Harvard Business School Press.

Murphy, K. R. and J. N. Cleveland (1995). *Understanding Performance Appraisal*. Thousand Oaks: Sage.

Ockenfels, A., D. Sliwka, and P. Werner (2010). Bonus Payments and Reference Point Violations. *IZA Discussion Paper No. 4795*.

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

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

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

Prendergast, C. J. and R. H. Topel (1993). Discretion and bias in performance evaluation. *European Economic Review 37*, 355–65.

Rynes, S., B. Gerhart, and L. Parks (2005). Personnel Psychology: Performance Evaluation and Pay for Performance. *Annual Review of Psychology 56*, 571–600.

Schleicher, D. J., R. A. Bull, and S. G. Green (2009). Rater reactions to forced distribution rating systems. *Journal of Management 35*, 899–927.

Schotter, A. and K. Weigelt (1992). Asymmetric tournaments, equal opportunity laws, and affirmative action: some experimental results. *Quarterly Journal of Economics 107*, 511–539.

Scullen, S. E., P. K. Bergey, and L. Aiman-Smith (2005). Forced distribution rating systems and the improvement of workforce potential: A baseline simulation. *Personnel Psychology 58*, 1–32.

5 Appendix

Dependent Variable:	Finished Blocks	Correct Blocks	False Blocks	False/Correct Blocks
	BaseBase vs. FdsFds			
	(1) Periods 1-16	(2) Periods 1-16	(3) Periods 1-16	(4) Periods 1-16
Fds	3.500**	2.833**	0.667	0.0082
	(1.558)	(1.161)	(0.712)	(0.020)
Group Ability	0.259***	0.255***	0.0039	-0.0011
	(0.059)	(0.043)	(0.026)	(0.001)
Constant	17.69***	14.09***	3.608***	0.204***
	(2.943)	(1.875)	(1.391)	(0.041)
Observations	512	512	512	512
Number of Groups	32	32	32	32
Wald Chi^2	1847.14	2761.22	138.57	150.84

Robust standard errors in parentheses, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Random effects regression (period dummies included)

Table A1: The Performance Effect of Forced Distribution

Dependent Variable:	Group Output	
	BaseBase	
	(1)	(2)
	Periods	Periods
	9-16	9-16
Item: "I gave bad grades to motivate the workers"	0.989*	
	(0.571)	
Item: "I gave bad grades to sanction the workers"		1.305**
		(0.599)
Group Ability	0.654***	0.639***
	(0.119)	(0.075)
Constant	34.49***	34.68***
	(5.526)	(3.536)
Observations	128	128
Number of Groups	16	16
Wald Chi^2	146.47	415.85

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

Random effects regression (period dummies included)

Items on a 7-point scale running from 1 "does not apply at all" to 7 "fully applies"

Table A2: The Impact of Deliberate Differentiation - Questionaire Items

		Game A					Game B				
		Pair I		Pair II			Pair I		Pair II		
		Payoffs (in €) for Player					Payoffs (in €) for Player				
	#	1	2	1	2	α_i	1	2	1	2	β_i
s	1	1.00	1.00	0.05	4.45	-0.19	5.00	0.00	0.00	0.00	1.00
w	2	1.00	1.00	0.71	4.39	-0.14	5.00	0.00	0.25	0.25	0.98
i	3	1.00	1.00	1.11	3.89	0.00	5.00	0.00	0.50	0.50	0.93
t	4	1.00	1.00	1.36	3.64	0.10	5.00	0.00	0.75	0.75	0.88
c	5	1.00	1.00	1.42	3.58	0.18	5.00	0.00	1.00	1.00	0.83
h	6	1.00	1.00	1.66	3.34	0.30	5.00	0.00	1.25	1.25	0.78
i	7	1.00	1.00	1.76	3.24	0.46	5.00	0.00	1.50.	1.50.	0.73
n	8	1.00	1.00	1.84	3.16	0.58	5.00	0.00	1.75	1.75	0.68
g	9	1.00	1.00	1.90	3.10	0.70	5.00	0.00	2.00	2.00	0.63
	10	1.00	1.00	1.93	3.07	0.79	5.00	0.00	2.25	2.25	0.58
p	11	1.00	1.00	1.96	3.04	0.85	5.00	0.00	2.50	2.50	0.53
o	12	1.00	1.00	2.03	2.97	1.00	5.00	0.00	2.75	2.75	0.48
i	13	1.00	1.00	2.07	2.93	1.18	5.00	0.00	3.00	3.00	0.43
n	14	1.00	1.00	2.09	2.91	1.30	5.00	0.00	3.25	3.25	0.38
t	15	1.00	1.00	2.12	2.88	1.41	5.00	0.00	3.50	3.50	0.33
	16	1.00	1.00	2.14	2.86	1.54	5.00	0.00	3.75	3.75	0.28
I	17	1.00	1.00	2.16	2.84	1.66	5.00	0.00	4.00	4.00	0.23
to	18	1.00	1.00	2.18	2.82	1.77	5.00	0.00	4.25	4.25	0.18
II	19	1.00	1.00	2.19	2.81	1.90	5.00	0.00	4.50	4.50	0.13
	20	1.00	1.00	2.21	2.79	2.02	5.00	0.00	4.75	4.75	0.08
	21	1.00	1.00	2.22	2.78	2.13	5.00	0.00	5.00	5.00	0.03
	22	1.00	1.00	2.50	2.50	2.18	5.00	0.00	5.25	5.25	0.00

Table A3: Eliciting Fehr/Schmidt parameters for inequity aversion using an adjusted version of the experimental procedure developed by Dannenberg et al. 2007. "#" indicates the unique switching point from pair I to pair II.

Dependent Variable:	Grade = 1	Grade = 4 or 5
	FdsBase vs. BaseBase	
	(1)	(2)
	Periods	Periods
	9-16	9-16
Output	0.422***	-0.339***
	(0.066)	(0.071)
FdsBase	-2.183***	1.395**
	(0.658)	(0.673)
Constant	-9.567***	3.043***
	(1.541)	(1.091)
Observations	493	493
Number of Subjects	93	93
Wald Chi^2	47.84	25.16

Standard errors in parentheses

*** p < 0.01, ** p < 0.05, * p < 0.1

Random effects probit regression (period dummies included)

Table A4: Ratings of Middle and Low Performers in the 2nd Part of the Experiment

Sample instructions for the first part of the experiment

First Part

This is the beginning of part one of the experiment. Please read the following instructions carefully. After having read the instructions you will find some test questions on your screen. The first part of the experiment is going to start as soon as all participants will have answered all the questions correctly.

Summary

The first part of the experiment consists of 8 rounds. Each round lasts two and a half minutes. In each round there are 4 participants per group. The group composition will be kept constant over the 8 rounds. No participant

will ever learn about the identity of any other participant in the group.

In this part of the experiment there are supervisors and workers. Out of the 4 participants per group one has the role of the supervisor and the other three are workers. The workers are denoted as "Worker A", "Worker B" or "Worker C". You will keep this name during the whole part.

Worker's Task

Each of the 8 rounds follows the same rules: the worker's task is identical to the task in the pre-round. She/he repeatedly has to identify the correct number of sevens in blocks of randomly generated numbers.

- Each block correctly solved is worth 2 points.
- Each wrong answer is worth -0.5 points, which means that if you state a wrong number of sevens there will be a penalty of half a point.

The number of correct and wrong answers results in the worker's total points of the round. The minimum number of points per round is zero which means that one cannot get a negative result.

As in the pre-round the worker can always press the "time-out button". If this button is used the worker's screen is locked for 20 seconds. During this time he cannot enter an answer. The time for the round keeps running during the time-out. So the worker loses 20 seconds per time-out since she/he cannot work on a block during this time. Please note that you cannot take a time-out during the last 20 seconds of a round.

Supervisor's Task

At the end of each round the supervisor gets to know the following for each worker in his group:

- The number of blocks correctly solved
- The number of wrong answers
- The resulting number of points

Then the supervisor rates the workers on a scale from 1 to 5, while 1 is the best (highest) and 5 is the worst (lowest) grade.

[Only FDS: Note: Each supervisor has to rate one of the workers with

"1" or "2", another one with "3" and one with "4" or "5" after each round.]

After the supervisor has completed her/his rating the workers get to know the following:

- The number of tasks correctly solved and number of wrong answers by herself/himself and the other workers in the group
- The resulting points
- The own rating (not those of the others)
- The own frequency of pushing the "timeout-button"
- The own payment for the round

Payment

Please note: Even though the amount is displayed after each round only one of the 8 rounds will actually be paid out. The payoff-relevant round will be publicly allotted at the end of the experiment. As the round will be randomly identified each of the eight rounds could be relevant for your payment which you will receive for the first part of the experiment.

Supervisor's Payment

The supervisor's payment is solely determined by the points achieved by his workers in the round. For each point achieved by a worker the supervisor gets 30 cents.

Worker's Payment

The worker's payment is determined by the rating assigned by the supervisor for the round:

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

For the grade "1" the worker would receive 10 Euros, for a "2" 7.50 Euros,

for a “3“ 5 Euros, for a “4“ 2.50 Euros and for a “5“ 0 Euro.

In addition to that the payment is determined by the frequency of pushing the „timeout-button“. Per usage of the “timeout-button” the worker gets 25 cents.

If there are any questions left please raise your hand. We will then come to your cabin.