Compensation, Performance and Selection - Empirical Studies on the Effectiveness of Incentive Schemes in Firms

Inauguraldissertation

zur

Erlangung des Doktorgrades

der

Wirtschafts- und Sozialwissenschaftlichen Fakultät

der

Universität zu Köln

2011

vorgelegt von

Diplom-Volkswirt Johannes Berger

aus

Marburg

Referent:	Professor Dr. Dirk Sliwka
Korreferent:	Professor Dr. Bernd Irlenbusch
Tag der Promotion:	16.12.2011

Für meine Eltern

Danksagung

Mein allererster Dank gebührt meinem Doktorvater Dirk Sliwka, dessen Geistreichtum und konstruktive Kritik mir sehr geholfen hat, Erkenntnisse immer wieder neu zu hinterfragen und meine Analysen weiter zu verbessern. Ich bin ihm überaus dankbar für seine loyale und kollegiale Art der Zusammenarbeit und seine großartige Führung während der gesamten Promotionszeit. Für die hilfreichen Kommentare sowie die Übernahme des Korreferates danke ich Bernd Irlenbusch sehr herzlich.

Von zentraler Bedeutung war für mich die Unterstützung meiner Lehrstuhlkollegen. Ganz besonders möchte ich mich bei den jetzigen und ehemaligen Mitstreitern Torsten Biemann, Kathrin Breuer, Anastasia Danilov, Christine Harbring, Björn Hartig, Claus Herbertz, Patrick Kampkötter, Felix Kölle, Petra Nieken und Nannan Zhou für ihre offene und unkomplizierte Zusammenarbeit bedanken. Ohne ihre große Diskussionsbereitschaft und ihre zahlreichen Denkanstöße wäre diese Arbeit nicht möglich gewesen. Darüberhinaus möchte ich mich bei unserer Sekretärin Beate Ommer sowie unseren studentischen und wissenschaftlichen Hilfskräften Isabella Cetnarowski, Paul Cibu-Deaconu, Alexander Creson, Behnud Djawadi, Saskia Günther, Andrea Hammermann, David Hirschfeld, Thorsten Hormesch, Katharina Laske, Michael Lüttjohann, Christiane Schäfer, Ursula Schuh, Kai Seifert, Julia Sohn, Andreas Staffelt und Timo Vogelsang für ihre große Hilfsbereitschaft bedanken.

Auch den (ehemaligen) Mitarbeiterinnen und Mitarbeitern vom Great Place to Work Institut, Frank Hauser, Johanna Kirsch und Karsten Schulte danke ich für die hervorragende Zusammenarbeit.

Besonders möchte ich auch Steffen Altman, Armin Falk und Matthias Wibral von der Uni Bonn danken, die mein Interesse für Verhaltensökonomie geweckt und mich in meinem Entschluss zu promovieren bestärkt haben. Vielen Dank auch an Jordi Brandts, Matthew Ellman, Julian Rohde und die anderen Mitarbeiter des Research Chair Antoni Serra Ramoneda in Barcelona, die mich während meines Auslandsaufenthaltes herzlich bei sich aufgenommen haben und mir fachlich und menschlich eine große Hilfe waren.

Ein ganz besonderer Dank geht an meine Eltern und meine Geschwister Gesine und Martin, die mich sowohl während des Studiums, als auch in der Promotionszeit in meinen Entscheidungen getragen und bestärkt haben. Danken möchte ich auch meinem guten Freund und mittlerweile Arbeitskollegen Julian Conrads. Unser permanente Austausch über die Diskrepanz zwischen ökonomischer Theorie und experimentalökonomischer Evidenz hat mich immer wieder neu in meinem Forschungsvorhaben bestärkt.

Ein großer Dank gilt auch all denjenigen ehemaligen Studienkollegen aus Trier, Indiana, Bonn und Barcelona, die dazu beigetragen haben, dass das Studentenleben auch nach Vorlesungsende noch abwechslungsreich verlief. Hier zu nennen sind vor allem Marc Niewalda, Johannes Stolz, Tobias Springer, Cory Leighty, Ina Karabasz, Javier Pinzon, Michael Muckle, Orsola Garofalo, Christina Rott und Max Weber.

Mein größter Dank geht allerdings an Barbara Dreyer, die mich mit großer Geduld, Ruhe und Selbstlosigkeit durch alle Höhen und Tiefen meiner Promotionszeit getragen hat. Ich bin ihr unendlich dankbar für ihr Gespür, mich in den richtigen Momenten von der Forschung abzulenken und an anderen Abenden über Papern, Emails oder Auswertungen alleine sitzen zu lassen. Ohne ihren starken Rückhalt während der letzten 3 Jahre wäre mir Vieles sehr viel schwerer gefallen.

Contents

1 Introduction **1**

2 Performance Appraisals and the Impact of Forced Distribution - an Experimental Investigation **10**

2.1 Introduction 10
2.2 Experimental Design 14
2.3 Results 19
2.3.1 Performance Effects of Forced Distribution 19
2.3.2 Differentiation and Productivity 27
2.3.3 What drives Rating Behavior? 35
2.3.4 Introducing or Abolishing a Forced Distribution? . . . 38
2.3.5 Forced Distribution and Costly Grades 42
2.3.6 Forced Distribution and Sabotage 44
2.4 Conclusion 47
2.5 Appendix 49

3 Heterogeneous Contestants and the Intensity of Tournaments - an Empirical Investigation **61**

3.1 Introduction 61
3.2 Related Literature 63
3.3 The Data 64
3.3.1 The Game of Handball 65
3.3.2 Heterogeneity Measures 65
3.3.3 Effort Measures 66
3.4 Results 68

3.4.1 The Impact of Heterogeneity on the Intensity of the Game . . . 69
3.4.2 The Impact of Heterogeneity on Favorites and Underdogs . . . 74
3.4.3 Testing our Measure of Game Intensity . . . 78
3.5 Conclusion . . . 82
3.6 Appendix . . . 83

4 Incentives and Cooperation in Firms - Field Evidence 89
4.1 Introduction . . . 89
4.2 Data and Hypotheses . . . 92
4.3 Results . . . 96
4.3.1 Team Incentives and Cooperation . . . 96
4.3.2 Incentives or Self-Selection? . . . 102
4.3.3 Team Incentives and Absenteeism . . . 106
4.4 Conclusion . . . 108
4.5 Appendix . . . 109

5 Gender Differences in Risk Preferences among Workers and Managers - Field Evidence from Germany 113
5.1 Introduction . . . 113
5.2 GSOEP: Data, Methods and Results . . . 115
5.3 GPTW: Data, Methods and Results . . . 123
5.4 Application: Performance-related Pay, Risk and Female Managers . . . 131
5.5 Conclusion . . . 133
5.6 Appendix . . . 135

Bibliography 139

List of Tables

2.1 Ratings and bonus payments 16
2.2 Overview of treatments in the core setting 18
2.3 The impact of forced distribution on productivity 24
2.4 The impact of rank and output on bonus payments 29
2.5 The impact of ratings on individual performance 32
2.6 The impact of deliberate differentiation on subsequent output 34
2.7 What drives rating behavior? 37
2.8 Effects of the introduction of a forced distribution 40
2.9 Introducing and abolishing forced distribution 41
2.10 Ratings, bonus payments and costs 42
2.11 Summary statistics of all treatments 49
2.12 The performance effect of forced distribution on different output measures 50
2.13 The impact of forced distribution depending on ability 51
2.14 The impact of forced distribution on productivity in the last 8 periods 52
2.15 The impact of ratings on timeouts 53
2.16 Eliciting social preferences - "#" indicates the unique switching point from pair I to pair II. 54
2.17 The Impact of forced distribution on productivity - treatment variations 55
2.18 The impact of deliberate differentiation on sabotage activity . 56

3.1 The effect of heterogeneity on 2-minute suspensions 71

3.2 The effect of heterogeneity on 2-minute suspensions of favorites and underdogs 77
3.3 Game intensity as a determinant for game outcomes 81
3.4 Calculating the winning probabilities and deriving a heterogeneity measure from sports betting odds 83
3.5 Descriptive statistics of key variables 84
3.6 The effect of heterogeneity on 2-minute suspensions - different heterogeneity measures 86
3.7 The effect of heterogeneity on 2-minute suspensions in each half 87
3.8 The relation between heterogeneity and 2-minute suspensions of favorites and underdogs 87
3.9 The effect of heterogeneity on the favorites' number of 2-minute suspensions at away games 88

4.1 Utilization of performance-related pay in the sample 94
4.2 Survey items approximating cooperation 95
4.3 Performance-related pay and cooperation among workers . . . 98
4.4 Performance-related pay, cooperation among workers and firm-size 99
4.5 Performance-related pay and cooperation among workers - subsample analysis 101
4.6 Performance-related pay and self-selection 104
4.7 Performance-related pay, self selection and cooperation among workers 105
4.8 Performance-related pay and absenteeism 107
4.9 Descriptive statistics of explanatory variables 110
4.10 Performance-related pay, HR policies and cooperation among workers 112

5.1 Gender-specific risk differences across hierarchical levels in the GSOEP - regressions 120
5.2 Gender-specific risk differences across hierarchical levels in the GSOEP - interacted regressions 122

5.3 Gender-specific risk differences across hierarchical levels in the GPTW - regressions 127
5.4 Gender-specific risk differences across hierarchical levels in the GPTW - interacted regressions 129
5.5 Share of women across hierarchical positions in Germany . . . 131
5.6 Share of female managers and performance-related pay 132
5.7 Descriptive statistics - GSOEP 136
5.8 Descriptive statistics - GPTW 137

List of Figures

2.1 Distribution of ratings across treatments 20
2.2 Distribution of group output across treatments 21
2.3 Group output over time across treatments 22
2.4 Distribution of ratings according to relative performance in the group 30
2.5 Group output over time across treatments when ratings are costly 43
2.6 Group output over time across treatments when sabotage is possible 46
2.7 Real-effort counting task in the experiment 50

3.1 The relation between heterogeneity (based on betting odds) and the number of 2-minute suspensions 69
3.2 The relation between heterogeneity (based on betting odds) and 2-minute suspensions of favorites and underdogs 75
3.3 The distribution of 2-minute suspensions in the sample 85

4.1 Utilization of performance-related pay across German industries 94
4.2 Structure of performance-related pay across German industries 94
4.3 The fraction of cooperative employees across German industries 111

5.1 Distribution of the general willingness to take risks in the GSOEP 2009 117
5.2 Gender-specific risk differences across hierarchical levels in the GSOEP - descriptives 118
5.3 Distribution of importance of job security in the GPTW 2006 125

5.4 Gender-specific risk differences across hierarchical levels in the GTPW - descriptives 125
5.5 Share of female employees across German industries in the GPTW data set 138

Chapter 1

Introduction

This thesis empirically investigates the effectiveness of incentives systems in firms and organizations. We are interested in how employees respond to different forms of monetary incentives. By the means of laboratory experiments and field data we analyze bonus systems based on subjective performance appraisals, tournament incentives and pay-for-performance incentive schemes. We study individual performance, cooperation, absenteeism and self-selection as potential employee responses evoked by these incentives systems. The interaction between individual preferences and compensation schemes is a second focus of this thesis. In particular, we provide evidence on the link between fairness concerns and evaluation behavior and the relevance of risk preferences for self-selection. Based on the empirical findings, we try to derive practical implications for the optimal design of incentives in firms. The following paragraphs briefly motivate the research questions of each chapter, summarize the main findings and explain how they relate to each other.

Since work effort is assumed to be costly to employees but beneficial to the firm, employment relationships comprise a natural conflict of interest. Resolving this conflict by appropriate incentive schemes is a central theme of personnel economics (see for instance Lazear and Shaw (2007)). If true effort was perfectly observable and thus contractible, incentives could easily overcome the moral hazard problem and realign interests by compensating employees for their effort costs. Since in practice performance rather than

effort is measurable, the design of optimal incentive plans is a delicate issue. Even performance is rarely fully captured by objective measures, which is why companies frequently rely on subjective performance evaluations to decide upon bonus payments. This, however, gives supervisors discretion over the distribution of ratings and thus the distribution of cash rewards in the respective department. As a result, supervisors can bias performance ratings according to individual preferences by, for instance, overstating true performances and neglecting performance differences among subordinates.

There are many reasons why supervisors may want to distort ratings. Psychology-based explanations refer to the mental costs associated with communicating negative feedback or the risk of rising frustration and envy in teams with higher pay dispersion. Alternatively, supervisors could try to signal superior leadership competencies by exaggerating low performers' ratings. Moreover, when performance measures are fraught with measurement error, supervisors may fear to make mistakes and prefer to pay everybody the same than to pay someone less who actually deserves more. Likewise, some supervisors simply want to avoid pay inequality among subordinates or between themselves and their subordinates for fairness reasons. Inflated and compressed ratings may thus be the result of generosity or inequity aversion. Indeed, several studies find the distribution of subjective performance evaluations to be heavily biased to the top and compressed to the middle of the rating scale. Since performance changes are no longer reflected in bonus changes, incentives to improve subsequent performance should be watered down (see Prendergast (1999) for a survey).

However, recent behavioral economic models may also predict the opposite. Reciprocity models (e.g. Rabin (1993), Dufwenberg and Kirchsteiger (2004) or Falk and Fischbacher (2006)) or the Fair-Wage-Effort Hypothesis (Akerlof and Yellen (1988), Akerlof and Yellen (1990)) suggest that inflated bonus payments could be perceived as kind gifts which employees may want to reciprocate by putting in extra effort in subsequent years. Also bonus differentiation may be perceived as unfair and -according to recent outcome-based models of inequity-aversion (e.g. Fehr and Schmidt (1999) and Bolton and Ockenfels (2000)) - cause utility decreases for inequity averse employees.

In response to the potential negative effects associated with biased ratings, companies have implemented rating systems that prevent lenient and compressed ratings. By assigning ranks or sorting fixed percentages of workers in pre-defined intervals of a performance distribution, supervisors are sometimes forced to distinguish between good and poor performance. Such systems, of which GE's "vitality curve" or the recently introduced forced ranking system at AIG are famous examples, remain controversially discussed. While some cherish the benefits of a competitive work environment, others stress the lack of acceptance among the workforce. Recent law suits initiated by employees who felt discriminated led some firms to abolish such systems again.[1]

Chapter 2 of this thesis investigates the determinants and incentive effects of biased performance ratings in more detail. We use a real-effort laboratory study in which a supervisor repeatedly evaluates the performance of three subordinates on a five-point scale. Subordinates are paid according to their individual rating while supervisors benefit from higher group performance. In the baseline treatment condition supervisors are unrestricted in their evaluation behavior and may inflate or compress ratings as much as they like. In another treatment supervisors must follow a pre-specified rating distribution, forcing them to differentiate between the top, middle and low performer in their group. The main result of the study is that under enforced differentiation workers are roughly 5-12% more productive over the course of the experiment. Importantly, the output increase does not come at the cost of lower output quality.

The key results remains robust, also when the supervisor has to carry some of the costs associated with employees' bonuses. However, a forced distribution seems to harm individual performance, when employees get the chance to sabotage each other. We observe a significantly higher amount of sabotage activity under the forced distribution system.

While most supervisors in the baseline treatment inflate and compress ratings to maximize subordinates' bonuses, those that deliberately differen-

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

tiate significantly improve group performance in the following periods. The study also reveals that rating biases can be explained by supervisors' concerns for altruism or inequity aversion.

When bonus differentiation is enforced, workers essentially compete for the largest bonus payment in the group. *Relative* performance becomes more important than *absolute* performance. The performance increase in the forced distribution treatment may thus be attributed to the willingness of subordinates to outperform their coworkers. To give subjects a realistic chance to affect their relative performance rank, we deliberately matched individuals with others of similar abilities into groups.

In reality, however, individuals that compete for a wage increase, a promotion or a contract may considerably differ with respect to ability, experience or skill. When ability differences are too large and evident, the inferior contestant may decide to save effort costs as his chances to win are comparably small. Anticipating such behavior, the superior contestant may decide to hold back effort as well. As a result, the overall effort level and the intensity of the tournament decreases. With regard to the study presented in chapter 1, one could infer that the productivity gains associated with forced distribution are smaller in heterogeneous work groups. While the effect of heterogeneity in tournaments is theoretically well understood (see for instance Lazear and Rosen (1981)), only recently empirical studies provide first evidence in line with this prediction.

In Chapter 3 we add to the emerging empirical literature on this question by analyzing professional sports data from the first German handball division. Statistics on sports contest often provide information on the ability or performance of tournament participants, measures that are usually not included in firm data sets but needed to test tournament theory. In our study we explore game-specific sports betting odds to estimate team abilities and collect data on the number of penalties committed by either team to measure defensive efforts and the overall intensity of the tournament. While betting odds allow us to construct a "market-efficient" measure for ability differences between the competing teams, our measure of effort needs more explanation: We assume that a team who decides to put forth a lot of effort has to play

very physical defense. Often high defensive efforts will successfully prevent goals by forcing the opponent to either turn the ball over, miss the goal or commit a time-penalty. In some instances, defensive efforts will be outside the legal norm and result in a penalty. In contrast to a game in which neither team tries, penalties should thus be more common in high intensity games.

In line with standard tournament theory, our data show that contests between heterogeneous teams are less intense as the number of committed penalties is lower than in games between teams of similar ability. Our analysis also reveals that while in theory both, the low and the top performer, should optimally reduce efforts when paired for an asymmetric contest, only the ex-ante dominant handball team plays less intensely. In our data, the underdog shows no significant reaction to the heterogeneity of the match up and tries to win "against all odds". Further sub-analysis also reveals that ex-ante ability differences are not only a good predictor of tournament intensity toward the beginning but also toward the end of the game. Irrespective the observed halftime score, larger differences in ex-ante winning probabilities are associated with less intense play in the second half. Finally, we are able to show that penalties may indeed serve as an effort measure as teams who commit more penalties (as a by-product of high effort) are more likely to win.

Since contests in sports and firms share essential characteristics, our study points out that promotion tournaments may not be an effective incentive instrument when the competing individuals or teams considerably differ with respect to ability or skill. In such instances firms may have to respond to ex-ante differences by handicapping the more able contestant at the beginning of the tournament or refraining from using tournament incentives at all.

Chapters 2 and 3 thus analyze two incentive schemes frequently applied when individual performance is either not objectively measurable or only measurable in relative terms. Both studies also highlight potential inefficiencies when individual incentives are used the wrong way. But companies may not only be interested in eliciting optimal individual efforts, especially if this comes at the expense of low cooperation. Nowadays a substantial amount of work is organized in teams (e.g. Lazear and Shaw (2007)). Teams are

usually installed when there are benefits from combining individual efforts. Due to, for instance, tasks interdependencies, the productivity of each individual's contribution may depend on the amount of team effort invested by other team members. For firms to benefit from teamwork, it is therefore essential that individuals are willing to cooperate. Incentivizing individual performance goals but not team performance can cause employees to allocate the bulk of their work time to individual rather than team assignments. This type of distortion has been analyzed by Holmström and Milgrom (1991) and is generally referred to as the multi-tasking problem. Even worse, when individual incentives are based on relative performance, as for instance in tournaments or under a forced distribution, employees may fully cease helping or even decide to sabotage work activities of their colleagues (see Lazear (1989) or Harbring and Irlenbusch (2011) and Chapter 1 for experimental evidence). To prevent this crowding out of cooperation, firms frequently rely on group incentives such as profit sharing schemes or rewards based on unit or department success.

From an economic perspective, the effect of group incentives on cooperation is unclear. On the one hand, individuals should cooperate more than in the absence of team incentives as they benefit from higher team output. On the other hand, team incentives generally bear the risk of free-riding since individual contributions to a team output are usually not observable. (Holmström (1982)). Especially when combined with individual incentives, employees are given an incentive to free-ride on their group members team efforts to save time and costs, better invested in meeting individual performance targets. Of course, especially in smaller teams, free-riding is unlikely to be tolerated by others and cooperation may be maintained via mutual monitoring or "peer-pressure" (e.g. Kandel and Lazear (1992)).

In chapter 4 we empirically investigate the relation between compensation schemes and cooperativeness in firms. We examine a representative sample of 305 German firms from the year 2006 that provides detailed information on the existence and strength of performance-related pay components, i.e. the amount of worker pay that depends on individual, team or firm performance. We map this information to roughly 36,000 employee survey

responses, measuring the level of cooperation among workers in these firms. Our main result is that 10 percentage points higher team incentives are associated with an 11% rise of employees who confirm that workers cooperate in their firm. Individual performance incentives or firm-level incentives do not relate to cooperation in the workforce. We also find that cooperation is in general higher in smaller teams and that the association between team incentives and cooperation is stronger in smaller companies. Since employees have fewer within-firm exit options and are more likely to interact with work colleagues in the future, free riding on team incentives may be less likely to occur. In addition, employees are also less frequently absent in firms with higher team incentives.

The results are robust to subsample analyses and the inclusion of several other control variables, capturing corporate culture and the overall level of job satisfaction among workers. The results also do not seem to be driven by cooperative workers self-selecting into firms that use team incentives. However, we do find that workers with preferences for helping are more likely to work in the health and social assistance industry and less likely to work in financial or business-related services.

In general, the question whether workers self-select into companies depending on the compensation scheme in place is a relevant one. Theoretically, performance-related pay should attract above average performing individuals because they should earn more than under a fixed wage contract. In contrast, below average performing employees should usually prefer contracts with lower variable pay components. The seminal case study by Lazear (2000) shows that selection indeed matters. Roughly half of the productivity increase that followed the switch from fixed to variable pay contracts could be explained by more productive workers self-selecting into the company.

Besides affecting the talent of employees, the introduction of performance-related pay may also cause individuals with particular personality attributes to self-select in or out of a company. Performance pay could, for instance, attract competitive, overconfident, or risk seeking individuals. Risk averse individuals, in contrast, should dislike wage uncertainty and prefer a fixed wage. While in principle the relation between risk attitudes and incentive

is straightforward, only recently Bellemare and Shearer (2010) and Grund and Sliwka (2010) provided convincing field evidence that risk preferences can indeed explain the likelihood of working under incentive plans (see also Dohmen and Falk (2011) for recent experimental evidence).

Knowing that incentive schemes affect self-selection with regard to risk attitudes, it is of course important to understand the determinants of risk attitudes to foresee the consequences of introducing or changing incentive plans. One of the most intensely studied determinants of risk aversion is gender. In particular, there is a general consensus among recent experimental studies that women are more risk averse than men (see e.g. Eckel and Grossman (2008b) and Charness and Gneezy (2010) for recent surveys). In a recent literature overview Croson and Gneezy (2009), however, point out that gender-specific differences in risk preferences do not extend to professionals and managerial employees. According to Croson and Gneezy (2009), one explanation could be that only employees who are willing to take risks select themselves into managerial positions and that after this pre-selection risk preferences are similarily distributed between women and men.

In the last chapter of this dissertation I investigate gender-specific differences in risk preferences among managerial and non-managerial employees in more detail. In particular, I study two large surveys that are representative for the German working population. At first, I analyze the 2009 wave of the German Socio-Economic Panel (GSOEP). The data set includes relatively detailed information on job hierarchy, a validated measure of general risk taking (Dohmen et al. (2011)) and two additional items measuring risk attitudes in job-specific contexts. In addition to the GSOEP, I also make use of the data set described in chapter 4. While this data set only provides a proxy for individual risk preferences, it includes observations from male and female employees from the same enterprise. This allows to control for unobserved company specific fixed-effects which is not possible in the GSOEP data.

The main result of the study is that in both data sets substantial and significant gender-specific differences in risk attitudes exist, not only among non-managerial employees but also on the managerial level. Moreover, the gender-specific differences do not systematically vary across levels. Indepen-

dent of gender, managers are more willing to take risks than employees on lower levels. Thus while it could be true that more risk-loving employees self-select into managerial positions, the effect is not stronger for women as conjectured by Croson and Gneezy (2009). The second data set also allows me to investigate the connection between incentive pay and risk preferences. In line with recent field evidence by Bellemare and Shearer (2010) and Grund and Sliwka (2010), I find that employees who receive performance-related pay are less risk averse than employees receiving a fixed wage. The result seems to be particularly robust for managers.

Chapter 2

Performance Appraisals and the Impact of Forced Distribution - an Experimental Investigation[1]

2.1 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.[2]

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

[1] This chapter is based upon Berger et al. (2010).

[2] 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).

not differentiate enough between high and low performers such that ratings tend to be compressed relative to the distribution of the true performance outcomes.[3] 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, the mechanisms behind these biases and the effects on performance have only rarely been analyzed empirically. 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. 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.[4]

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

[3]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).

[4]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. (2010), Kube et al. (2010), Bellemare and Shearer (2009) and Hennig-Schmidt et al. (2010).

discussed and in some firms even led to lawsuits as employees claimed to have been treated unfairly.[5]

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 within the organization is put at risk as 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 key reason for the lack of field evidence on the consequences of a forced distribution is that, even when a firm changes its performance appraisals 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 the performance consequences an objective measure of individual performance is necessary. But such objective measures are typically not available when subjective assessments are used.[6]

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 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 then asked to individually rate their performance on a five-point scale. The employees receive a bonus payment based on this performance rating. We

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

[6] 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.

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. Moreover, we study a setting in which supervisors share the costs of the bonuses paid out to the subordinates, as well as two additional treatments in which subordinates can sabotage each other.

Our key result is that worker productivity in our experiment is about 5-12% higher under a forced distribution system when there is no possibility to interfere with the colleagues' work. 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, knows about his relative performance and 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. In contrast, 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. The key results are robust in a situation in which it is costly for the supervisors to assign high bonuses.

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 (2011), Bol (2011), Kampkötter and Sliwka (2011) and Ahn et al. (2010) 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 2.3. 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. Moreover, we investigate the effect of past experience in a different rating setting on both, supervisor and the worker behavior. Finally, we analyze two additional experiments, one in which ratings are costly for the supervisors and one in which workers are able to sabotage each other. We discuss and conclude our results in the last section.

2.2 Experimental Design

We conduct a laboratory study investigating several different treatments. In all treatments subjects in the role of a "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 worker 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. In the following, we describe the structure of our core setting.

Ability Test

In an initial pre-round all subjects have to work on a real-effort task which 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 (see figure 2.7 for a screenshot of this task). This pre-round is independent of the main experiment and 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 of all, 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 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 each subject's performance is measured by the number of 'points' collected 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 pre-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 "timeout" button which locks the screen for 20 seconds during which subjects cannot work on any blocks. Each time the timeout button is pushed, the subject receives 8 cents. This timeout button is implemented to simulate potential further 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.[7]

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

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

to answer several test questions on the screen to make sure they have fully understood the procedures and the payoff calculations.[8] This first part of the experiment consists of eight periods, each lasting for 2.5 minutes. Each participant is assigned to a group 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 timeout button, blocking the screen for 20 seconds for which they receive 25 cents on their private account. After each round, each worker learns her total number of points, the number of correct and false answers and the number of timeouts chosen. Moreover, each worker is also informed about the number of points and correct and false answers of the other two workers in her 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 2.1: Ratings and bonus payments

Each rating is associated with a bonus payment for the worker (see table 2.1), ranging from 10 € for the highest rating "1" to 0 € for the worst rating of "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 who evaluate the performance of employees are not residual claimants but are themselves salaried employees and, hence, higher bonus payments to subordinates do not lower their own income. We

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

will later on vary this and investigate treatments in which supervisors bear costs for higher bonus payments.

The round payoff for the worker is the sum of her bonus payment and the payoff from pushing the timeout 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 her rating, the number of timeouts and her resulting payoff. The worker does not learn about the other workers' ratings in her 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.[9] 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

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*), however, supervisors have to give one

[9] In one of our extensions we informed the subjects about the matching procedure and added survey questions in the end of the experiment to investigate potential effects of the procedure. However, we did not find evidence that this affected the way in which evaluations were conducted or the reactions to the evaluations.

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 this setting (see table 2.2).

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

Table 2.2: Overview of treatments in the core setting

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 told 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.[10]

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

[10]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.

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 core setting of the experiment consists of 8 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 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)).

2.3 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 2.3.4 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 experiment, one in which awarding bonuses is costly for the supervisors and one in which workers can sabotage each other.

2.3.1 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 have thus 32 strictly independent group observations.[11]

[11]Note that *BaseBase* and *BaseFds* on the one hand and *FdsFds* and *FdsBase* on the other are perfectly identical in the first part as participants only learn the rules of the second part after the end of the first one.

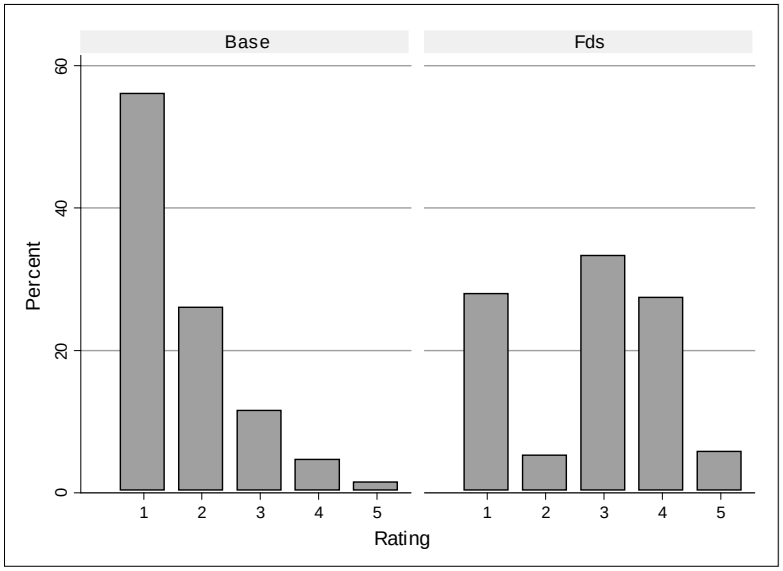

Figure 2.1: Distribution of ratings across treatments

Figure 2.1 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), 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 in most real-world organizations, supervisors in the experiment do not have to bear the direct costs of higher bonus payments.[12] 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

[12] A setting in which higher ratings are costly for the supervisors is studied in section 2.3.5.

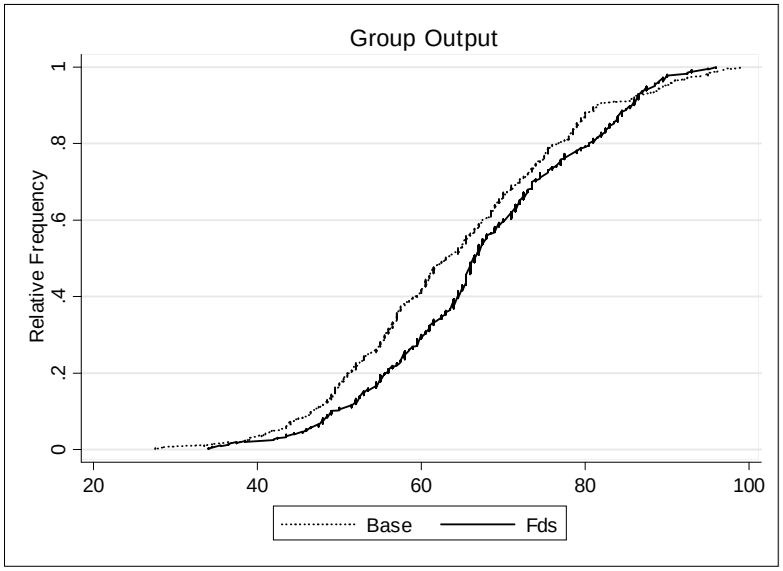

Figure 2.2: Distribution of group output across treatments

of freedom left by the system, the supervisors in *Fds* 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 2.1.

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 as compared to the forced distribution treatment. Hence, participants in the role of employees 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. As already laid out in the introduction, numerous gift-exchange experiments have provided evidence for the fair wage-effort hypothesis, positing that higher wage payments may lead to higher efforts.

Figure 2.2 plots the distribution of group output in both treatments. The figure indicates that performance is indeed higher under the forced distribution. Group performance increases on average by about 5% and the difference raises to almost 9% when we analyze the second parts of the *BaseBase* and *FdsFds* treatments.

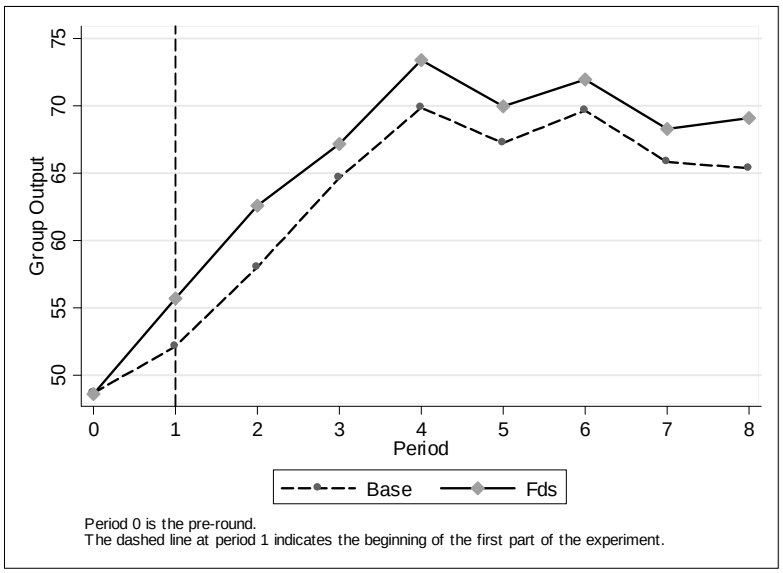

Figure 2.3: Group output over time across treatments

Taking a closer look at the evolution of work performance over time in figure 2.3, we see that while performance is identical prior to the treatment intervention (in the pre-round), average performance is substantially higher across all periods of the experiment.[13]

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. Due to the matching procedure we control for the number of points achieved by the group or the individual in the pre-round (period 0).[14]

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 pre-round performance using

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

[14]Note that the matching of participants into homogeneous groups resulted in a very low standard deviation of outputs within the majority of groups. In the pre-round, the average standard deviation of worker output amounts to 0.71 output points. In only 6 out of 64 groups the standard deviation exceeds 1.5 output units.

only one data point per group.[15] As 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 which include periods dummies to control for the general time trend observed in figure 2.3. In a thirdalternative 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 2.3. 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 dependent variable.

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

Dependent Variable:	Output			Log Output		
	Base vs. Fds (periods 1-8)			Base vs. Fds (periods 1-8)		
	OLS	RE (Groups)	RE (Individuals)	OLS	RE (Groups)	RE (Individuals)
	(1)	(2)	(3)	(4)	(5)	(6)
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)
Pre-Round 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 Chi^2	0.68	741.10	723.90	0.69	532.75	523.67

Robust standard errors in parentheses (in (3) and (6) clustered on group_id), *** p < 0.01, ** p < 0.05, * p < 0.1

(1 & 3) Ordinary least squares regression on collapsed average group output (one observation per group)

(2 & 4) Random effects regression on periodic group output

(3 & 6) Random effects regression on periodic individual output

(2-3 & 5-6) Period dummies included

Table 2.3: The impact of forced distribution on productivity

Column 1 shows that the forced distribution indeed significantly increases group performance by roughly three output units. This corresponds to a 5.6% increase in group performance as displayed in model (4).[16] The coefficients obtained in the random effects models parallel these results. Furthermore, in all specifications pre-round performance is strongly correlated with actual performance in the experiment. The estimate in column 1 suggests that groups that solved one block more in the pre-round on average solved half a block more 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 2.12 in the appendix), we find that under forced distribution subjects count and solve more blocks correctly while making only slightly and insignificantly more mistakes.

To provide an even more conservative test without any distributional assumptions, we additionally apply the following non-parametric procedure: Due to our matching mechanism, groups within a treatment are, by definition, not drawn from the same population, but groups of the same rank with respect to the pre-round performance are directly comparable across treatments. We thus rank group observations in each treatment according to their pre-round performance from 1 to 32 and calculate the output difference of each group with its counterfactual in the other treatment. E.g. the average group output of the 8^{th} able group in the *Fds* condition is compared with the 8^{th} 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 is higher in the *Fds* groups. This difference is statistically significant in a one-sided binominal test ($p = 0.055$).[17]

In principle, our experimental design allows two explanations for why

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

[17] Applying the same test to test for differences in pre-round performance we see that (i) in 6 out of 32 comparison groups output was exactly the same and (ii) of the remaining 26 output is higher for 13 groups in *Fds* and 13 in *Base*. Hence, randomization performed very well such that ability is equally distributed across the two treatment groups (see also figure 2.3).

productivity increases under *Fds*. The observed treatment difference in performance may be the result of subjects working harder, i.e. they solve more blocks in a given amount of time, or taking less timeouts. Investigating the choice of timeouts, we find that the timeout option was rarely chosen in the two core treatments. On average, only 0.7 timeouts per group were taken in each round. Furthermore, there is no systematic difference in timeout usage across treatments.[18] If we either control for the number of timeouts in the regression or exclude group observations in which timeouts were taken, the treatment effect becomes stronger.

We also investigate whether the incentive effect of forced distribution is stronger among low or high talented groups. Table 2.13 extends our standard regression by an interaction term *Fds* x Pre-Round Group Output. The substantially larger and highly significant *Fds* coefficient and the negative interaction term reveals that forced distribution is particularly effective among low performing groups.[19]

Finally, we explore the performance effect of *Fds* in the second 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 2.14 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 non-parametric procedure ($p = 0.038$, one-sided Binomial test).[20]

[18] In the first eight periods, timeouts are slightly more frequent in the *Fds* condition but the difference is not statistically significant. In the last 8 periods of the treatments *BaseBase* and *FdsFds*, timeouts are less frequently used under forced distribution. In periods 9-16 of *Fds*, in only 8 out of 128 group observations (period x group) at least one timeout was observed compared to 48 out of 128 in the baseline treatment. However, this difference is also not significant, neither in regressions, nor in non-parametric tests.

[19] However, the key results are qualitatively robust when we drop the four lowest groups or when we drop the 10% highest and 10% lowest performing groups.

[20] When only considering the *BaseBase* and *FdsFds* treatment, there is a significant difference in the pre-round 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.

2.3.2 Differentiation 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 as 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 as they are not forced to differentiate. In contrast, variations in absolute performance may affect grading 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 of a worker as independent variables.[21] To illustrate treatment differences, we include interaction terms with a dummy variable for the forced distribution treatment.

The results are reported in table 2.4. Note that the relative rank matters in both treatments but does so to a much larger extent under forced distribution as indicated by the substantially larger rank coefficients in column (2) and the significant interactions of rank and *Fds* in columns (3-4). Interestingly, while within-rank variation in output is rewarded in both treatments, these rewards are stronger in the baseline treatment. For a given rank, output and bonus are more strongly (positively) correlated than under forced distribution. This is indicated by the substantially smaller output coefficient in column (2) and the significant negative interaction of output and *Fds* in columns (3-4). But, apparently, competing for ranks generates stronger incentives in the forced distribution treatment as shown by the positive interaction terms of ranks and *Fds*. The competition for ranks indeed induces a 'tournament' among the agents. As the literature on tournaments - starting with Lazear and Rosen (1981) - has pointed out, tournament competition

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

can indeed be a powerful incentive instrument.[22] However, it is interesting to note that supervisors in the baseline setting could have also implemented such a tournament but apparently did not condition on relative rank sufficiently to induce similar high powered incentives.

It is furthermore interesting to note that for a given output agents obtain a higher bonus the lower their performance in the pre-round. This is indicated by the negative coefficients for pre-round group output. Since our matching procedure produces homogenous groups, agents with higher pre-round performance are grouped together. Hence, a higher pre-round performance increases the reference level relative to which supervisors compare individual output. It may therefore be harder to obtain a high bonus in a stronger group.[23]

Figure 2.4 shows the distribution of grades for the top, middle, and low performers in the first eight periods across both treatment conditions.[24] In the forced distribution treatment 91% of the participants with the highest rank receive a "1" or a "2" and 88% with the lowest rank a "4" or a "5". In contrast, about 60% of the worst performers still receive a "1" or a "2" in the baseline treatment. Hence, 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 2.5 reports results from a random effects regression with individual output in $t+1$ as the dependent and dummy variables for the grade assigned in period t as independent variables. 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

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

[23]When we add dummies for the group rank, the effect of lower pre-round performance disappears and is instead captured by the significant group dummies. But all other regression coefficients and significance levels remain very similar. Note that group rank, worker and supervisor ability are highly correlated which makes it hard to disentagle the influences of each variable. Further analyses, however, suggests that rating behavior (e.g. the rating differentiation) does not depend on the ability of the supervisor.

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

Dependent Variable:	Individual Bonus			
	Base	Fds	Base vs. Fds	BaseBase vs. FdsFds
	Periods 1-8	Periods 1-8	Periods 1-8	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)
Pre-Round 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 Chi^2	468.10	1762.93	1855.37	3382.97

Robust standard errors in parentheses (clustered on group_id)

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

Random effects regression (period dummies included), reference category: rank 3

Table 2.4: The impact of rank and output on bonus payments

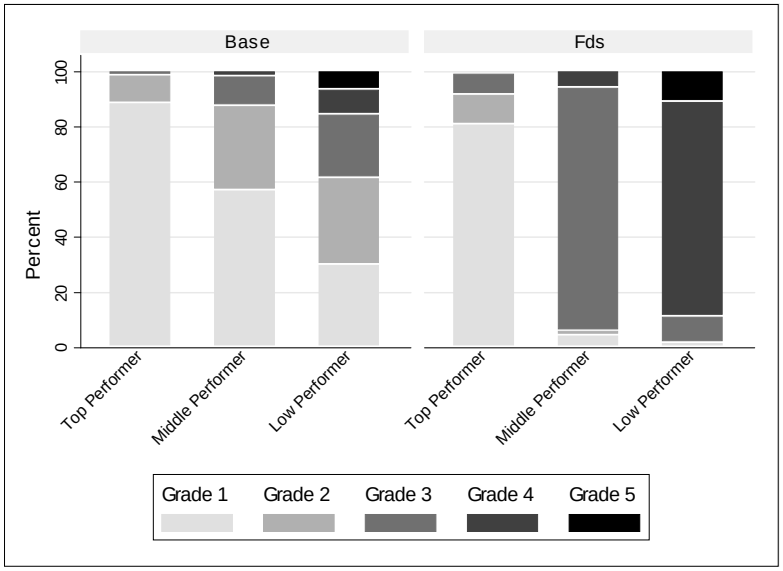

Figure 2.4: Distribution of ratings according to relative performance in the group

performers in each period. Since grade 5 is rarely observed we pool grades 4 and 5.

We indeed observe significant immediate reactions in those cases where the grade obtained is particularly informative about a supervisor's grading policy: Middle and low performers substantially increase their outputs after 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 top grade reduce their efforts which supports the view that lenient and undifferentiated ratings indeed undermine performance incentives.[25] When the forced distribution is in place, subjects know the rating policy as 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

[25]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.

directly infer that worse performing coworkers have obtained better grades and that high performance is not rewarded.

Dependent Variable:	Individual Output$_{t+1}$					
	Base (periods 1-8)			Fds (periods 1-8)		
	All Workers	Top	Middle/Low	All Workers	Top	Middle/Low
	(1)	(2)	(3)	(4)	(5)	(6)
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)
Pre-Round 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 Chi^2	1150.92	1133.28	585.71	587.55	233.88	180.78

Robust standard errors in parentheses (clustered on group_id)

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

Random effects regression (period dummies included), reference category: Grade=1$_t$

Table 2.5: The impact of ratings on individual performance

We also study timeouts as a potential reaction to ratings. Arguable, taking a timeout is an even simpler and less ambiguous measure of discontent or a lack of motivation. Table 2.15 therefore explains the sum of timeouts taken by an individual in period $t+1$ by the rating received in period t. Indeed, we find that agents reacted with their timeout choices to the grading behavior. The pattern in which they do parallels the relation of grades and output presented in table 2.5 (but of course the other way around). While we observe that in general receiving a "3" instead of a "1" decreases the likelihood of observing a timeout for that given worker in the next period, the effects of grades again depend on the relative performance rank of the individual. While top performers (column 1) are significantly more likely to take a timeout in response to a "2" instead of a "1", we find that giving worse grades to the middle and low performers has positive productivity effects (less timeouts).

These results suggest that supervisors will induce higher performance in subsequent rounds by differentiating in their ratings. 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 2.6. No differentiation, i.e. cases in which each worker receives the same rating, serves as our reference category.[26] The results suggest that extending the range of applied ratings from 0 to 1 in the first eight periods increases subsequent productivity on average by 4 1/2 points (6%) in the first part. Extending the range of grades from 0 to 2 also has a significant positive effect on subsequent performance. This effect seems to be larger in the second part of the experiment.[27]

[26] In 24% of all rounds in *Base* the supervisor assigned all workers the best rating "1" and in 25% of all rounds she/he assigned the same rating to all three participants. In the second part of *BaseBase* the percentages rise to 29% and 31% respectively.

[27] Note that an observed span of grades larger than 2 occured in only 31 out of 256 rating decisions in the first part of *Base*. Similar to the previous regression the results for large spans are mixed. While it seems to improve performance in the last 8 periods, it has no significant effect on subsequent output in the first part of the experiment. One potential explanation could be that some workers did not work at all after receiving such a

Dependent Variable:	Group Output$_{t+1}$	
	Base (periods 1-8)	BaseBase (periods 9-16)
	(1)	(2)
Span of Grades=1$_t$	4.416***	2.737
	(1.045)	(1.709)
Span of Grades=2$_t$	2.681**	6.367***
	(1.045)	(1.413)
Span of Grades=3 or 4$_t$	2.338	6.097**
	(2.685)	(2.749)
Group Output$_t$	0.604***	0.331***
	(0.0962)	(0.0870)
SD of Output$_t$	0.0637	-0.0665
	(0.188)	(0.231)
Pre-Round Group Output	0.257***	0.511***
	(0.0786)	(0.118)
Constant	10.45***	20.21***
	(2.782)	(3.330)
Observations	224	112
Number of Groups	32	16
Wald Chi^2	1073.37	4289.67

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

Random effects regression (period dummies included)

Reference category: span of grades=0$_t$

Table 2.6: The impact of deliberate differentiation on subsequent output

Additional evidence for positive effects of deliberate differentiation can be derived from our post-experimental questionnaire. As already mentioned above, we asked subjects in the role of the supervisors about their rating behavior in both parts of the experiment. The items[28] "*I assigned bad ratings to motivate the workers*" and "*I assigned bad ratings to sanction the workers*" are both positively correlated with a higher group output in the second part of the experiment (significant at the 10% and 5%-level). Moreover, these self reported measures of differentiation are highly correlated with actual differentiation in the second part of the experiment (e.g. span of grades), even after controlling for group output.[29]

2.3.3 What drives Rating Behavior?

As the personnel psychology literature[30] has already stressed, the personality of the rater affects 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 the way in which performance ratings are assigned. 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), 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 for maximizing the total surplus of all participants to some extent) or are altruistic and therefore directly care for the payoffs of others and thus should assign better grades that

low grade. Due to the increase in noise, the positive coefficient is not significant anymore.

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

[29] Regression results available upon request.

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

lead to higher bonuses.

To investigate these drivers we apply an adapted version of an incentivized experimental procedure introduced by Blanco et al. (2010) and modified by Dannenberg et al. (2007). It consists of simple choice experiments in which participants have to choose between pairs of payoff tuples, specifying a payment to themselves and to some randomly drawn other subject. In the first set of choices ("Game A", see table 2.16 in the appendix) 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 ("Game B") 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 which 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*.[31]

We now expect that both the altruistic and the equity oriented types assign better grades. But while the altruistic types should do that unconditionally, we should expect that equity oriented types make a stronger connection between performance and the assigned ratings. We do not expect that

[31] The relevant switching points are #3 for game A and #21 for game B. As stressed by Blanco et al. (2010) and Dannenberg et al. (2007) these games can be used to infer α and β in a Fehr and Schmidt (1999) type utility function. It is interesting to note that this procedure typially gives negative estimates of α for a non negligible fraction of subjects, i.e. those who are willing to sacrifice own payoff to increase overall surplus even though they are worse off than their counterpart. We classify these subjects as altruists.

As laid out by Blanco et al. (2010) (footnote 20 on p. 30) the Fehr Schmidt utility function also captures surplus concerns when allowing for negative values of α.

envious types' rating behavior differs from selfish ones as the supervisors are typically better off than workers.

Dependent Variable:	Group Bonus		Span of Grades	
	Base	Fds	Base	Fds
	(Periods 1-16)		(Periods 1-16)	
	(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)
Pre-Round 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 Chi^2	261.53	224.25	393.70	297.09

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

Random effects regression (period dummies included)

Reference category: 1=Selfish supervisors

Table 2.7: What drives rating behavior?

Table 2.7 now reports regression results with the total of bonus payments awarded to the group or the span of grades as dependent variables and dummy variables for the different types as independent variables (reference group are the selfish supervisors). As expected, we observe the supervisor's type indeed matters in the baseline setting. Altruistic types award the high-

est 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. Since supervisor's earn substantially more than workers, envious supervisors do not rate differently than selfish types.[32] 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 2.4 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.[33]

2.3.4 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. Because we have to take 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 above, we should expect forced distribution to increase performance in the second part of *BaseFds*. 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 2.8. However, a surprising pattern emerges when we compare the effects per period as shown in column (2). While performance first increases by about 5 points in period 9 and stays at this level in period 10, it drops to roughly 2-3 points below the baseline level in the last 6 periods. Hence, participants are ap-

[32] Social preferences do not explain rating behavior under forced distribution. Most likely, the rating scheme does not allow enough variation for ratings to be affected by individual preferences.

[33] Regression results available upon request.

parently initially motivated 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. In contrast to a setting in which a forced distribution is present from the outset, participants now have a different reference standard as they have experienced more favorable ratings in the past. This may lead to a decrease in motivation under *Fds*. 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 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 8 periods of *BaseFds* to the treatment in which the forced distribution has been used throughout the experiment (*FdsFds*). But as column (1) of table 2.9 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 payment loss apparently seems to counteract the positive forces of increased differentiation. The highly significantly difference in timeouts, displayed in column (2), 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) indicates that groups in which *Fds* has been abolished are roughly 7% more productive than workers in *BaseBase.* 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 feel more inclined to reciprocate this relative increase in bonus payments. Yet, another factor driving this result is that supervisors keep up

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)
Pre-Round 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 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 2.8: Effects of the introduction of a forced distribution

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* (for a given output). 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 the second part of *FdsBase* than in *BaseBase.* Also, as indicated by the negative *FdsBase* dummy in column 1 of table 2.7, ratings are on average lower than under *BaseBase*. 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		Group Timeouts	
	BaseFds vs. FdsFds	BaseBase vs. FdsBase	BaseFds vs. FdsFds	BaseBase vs. FdsBase
	Periods 9-16		Periods 9-16	
	(1)	(2)	(3)	(4)
BaseFds	-5.763*		0.735***	
	(2.994)		(0.268)	
FdsBase		4.591*		-0.0055
		(2.363)		(0.015)
Pre-Round 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 Chi^2	318.06	57.95	19.93	11.07

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

Table 2.9: 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 experiment. Especially workers in *BaseFds* feel that their effort paid off to a greater

extent during the baseline setting. They also state that the supervisor's behavior is more fair and that she is more capable of giving appropriate ratings in the absence of a forced distribution. The supervisors also express some dissatisfaction towards the forced distribution as, for instance, they perceive 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).

2.3.5 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. However, 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. Table 2.10 summarizes these costs. 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.[34]

Rating	Bonus Worker	Supervisor Costs
1	10.00 €	5.00 €
2	7.50 €	3.75 €
3	5.00 €	2.50 €
4	2.50 €	1.25 €
5	0.00 €	0.00 €

Table 2.10: Ratings, bonus payments and costs

[34]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 pre-round. We additionally extended the post-experimental questionnaire to check for potential effects of this procedure but did not find evidence that this affected participants' behavior.

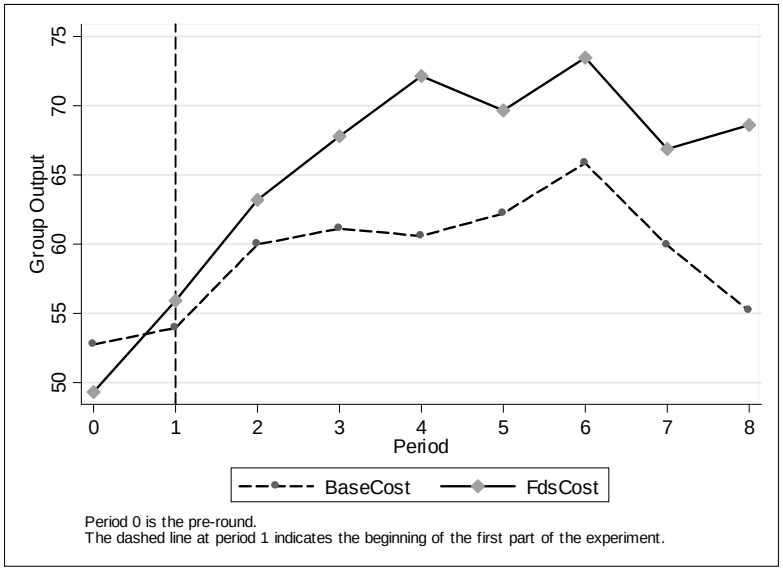

Figure 2.5: Group output over time across treatments when ratings are costly

Figure 2.5 shows the average output over time in the first part of the new treatments.[35] The results are qualitatively surprisingly similar to our earlier results and the effect of a forced distribution seems to have an even stronger impact on performance.

Average group output is 59.6 in *BaseCost* and 67.2 in *FdsCost* and even though groups in *BaseCost* are on average slightly more productive in the pre-round, performance is already higher in early periods of *FdsCost* and increases over time. As the regression results in column 3 of table 2.17 show, the performance difference amounts to 9.4 output units or 12% and is significant at the 5% level. Again this difference is also confirmed by the non-parametric testing procedure laid out in the above ($p = 0.059$, one-sided binominal test). This is the case even though in the pre-round groups in the *FdsCost* treatment are (weakly) significantly less productive than groups in the *BaseCost* treatment.

We also studied agents' behavior in a second part where participants worked for another 8 periods under the same rules. Interestingly, the treat-

[35] One group in *BaseCost* had to be dropped due to a technical problem with the experimental software.

ment difference is no longer significant in all periods. While the treatment coefficient is still substantial (6.99), the standard error is now much higher, indicating that output is noisier in the second part. One part of the reason is a sharp performance increase in period 9 of *BaseCost*. Here, the average performance in the baseline treatment even exceeds the performance under the forced distribution. A potential explanation is the following: In the baseline treatment there is a considerable endgame effect in period 8 as apparently workers anticipated low bonus payments in the last period and decided to put in less effort (see figure 2.7). After the unexpected restart, workers (i) were more rested and (ii) had an incentive to signal their willingness to work as the game continued for another 8 periods. Indeed, the number of group timeouts taken dropped from more than 5 in period 8 to less than 1 in period 9. This effect is absent under forced distribution. Finally, when only considering the last two periods of the second part, the treatment difference is significant again. Under the forced distribution workers know that even in the last period one of the agents must receive a high bonus. This avoids the endgame effect present in the baseline treatment when bonuses are costly.

2.3.6 Forced Distribution and Sabotage

The previous chapters 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 uncooperatively, deny help or even sabotage coworkers. Examples for such behavior could be withholding viable information or, in the extreme, deleting files on computers, stealing others' equipment or the like. It is crucial to understand that the effectiveness of incentives depends on the environment they are embedded in. Indeed, the literature on tournaments has stressed that tournament competition can create incentives, not only for productive work but also to sabotage each other (Lazear (1989) see Harbring and Irlenbusch (2011) for experimental studies on this issue).

With regard to systems of forced distribution Prendergast claims: *"Forced rankings also increase competition for merit pay, which is counterproductive in environments where cooperation is important to production"* (Prendergast and Topel (1993), p. 362).

We test this conjecture with a simple treatment variation of our current experimental setup. In addition to counting numbers and taking timeouts, subjects are explicitly given the opportunity to block a coworker's screen for 20 seconds such that the fellow worker can not work or take timeouts. This "sabotage option" is costly as the choice of blocking somebody else's screen blocks the own screen for three seconds, modeling the fact that sabotage also incurs some costs for the workers. There is no restriction on the frequency of sabotage, i.e. subjects can block other subjects as often as they like.[36] After being blocked for 20 seconds, it is ensured that subjects can not be sabotaged again within the next 5 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 8 periods each, keeping the two treatment conditions baseline (*BaseSabo*) and forced distribution (*FdsSabo*) unchanged in both parts.

The key hypothesis is that forced distribution should lead to higher sabotage activities as workers compete for the high ratings and can improve their position by harming coworkers. Together with our prior results we therefore conjecture that a trade-off exists as the forced distribution may increase incentives but may also induce wasteful sabotage activities.

Indeed, we find that subjects use the sabotage option twice as often under the forced distribution (about 8 times per group and period) than under baseline. Moreover, this difference leads to strongly detrimental consequences for overall group performance. As a result, average group performance under the forced distribution is as low as 33.3 which is 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 column 5 & 6 in table 2.17. Figure 2.6 depicts the performance over time across the

[36] However, they are told that there is no effect if the subject's screen they intended to sabotage is already blocked or if the subject is in a timeout.

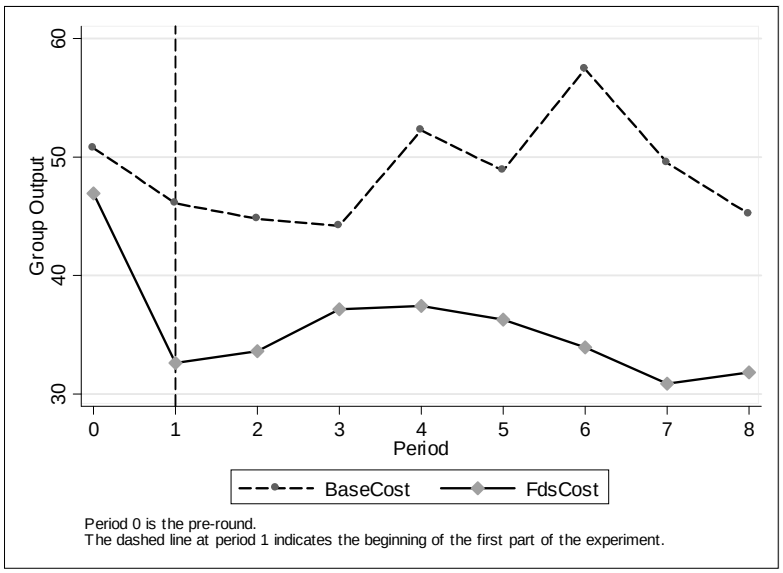

Figure 2.6: Group output over time across treatments when sabotage is possible

two settings and suggests that the performance difference even increases over time.

The treatment difference is also robust when we again apply the non-parametric test to compare the differences across groups of the same rank with respect to pre-round performance from both treatments ($p < 0.01$, one-sided binominal test).

It is furthermore interesting to note that higher degrees of differentiation also lead to more sabotage activity within baseline treatment alone and thus lower performance in subsequent periods (see table 2.18). Hence, more differentiation indeed sets incentives to outperform coworkers and the easiest way to do this 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 as can be seen in column 5 & 6 of table 2.17.

2.4 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 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. But even within the baseline setting those supervisors who choose less lenient and more differentiated ratings attain a higher performance.

Moreover, we analyze the supervisor's social preferences as potential drivers of rating behavior. We find that social preferences have a substantial impact on rating behavior in the baseline setting. Particularly, altruistic supervisors (as measured by simple choice experiments) tend to give higher bonuses while equity oriented supervisors choose significantly less differentiated ratings.

However, our results also indicate potential problems of using a forced distribution. First of all, it may be problematic to set up a forced distribution when employees have experienced a more "liberal" system of performance evaluations before. Most importantly, we find that introducing forced distribution into an existing appraisal system leads to a short-term performance increase, followed by a rather sharp drop in performance. Apparently, while 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. In contrast, some experience with the forced distribution in the beginning demonstrates supervisors the benefits of differentiation as they tend to differentiate more and are able to maintain a higher performance when forced distribution is abolished again.

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", e.g.

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.

In additional treatments we extended our experimental set up by allowing workers to temporarily prevent their coworkers from working on the task. The interesting result of these additional treatments is that sabotage activities occur much more frequently when workers compete for higher bonuses under forced distribution. This has detrimental consequences for overall group performance. 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.

Our study, thus, sheds some light on the prevalent problem of subjective performance evaluations in organizations and adds some empirical findings to the discussion on the effectiveness of forced distribution systems. Of course, there are still many further research questions. For example, it would be interesting to study the robustness of our results for different and more complex tasks or in settings where participants know each other well or can communicate with each other.

2.5 Appendix

Variables	Pre-Round Group Output	Group Output		Group Timeout		Group Rating		Number of Groups
Periods		1-8	9-16	1-8	9-16	1-8	9-16	
Base	48.67	64.11	71.45	0.54	0.71	1.70	1.75	32
Fds	48.63	67.28	73.00	0.84	0.60	2.78	2.74	32
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 2.11: Summary statistics of all treatments

Figure 2.7: Real-effort counting task in the experiment

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)
Pre-Round 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 Chi^2	626.24	756.05	13.63	19.53

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

Random effects regression (period dummies included)

Table 2.12: The performance effect of forced distribution on different output measures

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)
Pre-Round Group Output	0.635***	0.730***
	(0.0610)	(0.0974)
Fds × Pre-Round 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 Chi^2	768.71	186.13

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

Table 2.13: The impact of forced distribution depending on ability

Dependent Variable:	Output			Log Output		
	BaseBase vs. FdsFds (period 9-16)			BaseBase vs. FdsFds (period 9-16)		
	OLS	RE (Groups)	RE (Individuals)	OLS	RE (Groups)	RE (Individuals)
	(1)	(2)	(3)	(4)	(5)	(6)
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)
Pre-Round 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 Chi^2	0.66	179.43	181.33	0.64	192.19	184.69

Robust standard errors in parentheses (in (3) and (6) clustered on group_id), *** p < 0.01, ** p < 0.05, * p < 0.1

(1 & 3) Ordinary least squares regression on collapsed average group output (one observation per group)

(2 & 4) Random effects regression on periodic group output

(3 & 6) Random effects regression on periodic individual output

(2-3 & 4-5) Period dummies included

Table 2.14: The impact of forced distribution on productivity in the last 8 periods

Dependent Variable:	Individual Timeout$_{t+1}$					
	Base (periods 1-8)			BaseBase (periods 9-16)		
	All Workers	Top	Middle/Low	All Workers	Top	Middle/Low
	(1)	(2)	(3)	(4)	(5)	(6)
Grade=2$_t$	-0.107*	0.00347	-0.101*	-0.0828	-0.0159	-0.330***
	(0.0591)	(0.105)	(0.0604)	(0.0628)	(0.122)	(0.102)
Grade=3$_t$	-0.168*	0.0985	-0.226**	-0.154*	-0.199	-0.492***
	(0.0883)	(0.241)	(0.104)	(0.0933)	(0.159)	(0.142)
Grade=4 or 5$_t$	0.244	-0.318	0.155	0.0970	-	-0.325
	(0.316)	(0.217)	(0.303)	(0.341)	-	(0.293)
Output$_t$	-0.0481**	-0.00467	-0.0702**	-0.00477	0.0255**	-0.0676***
	(0.0239)	(0.00995)	(0.0305)	(0.0125)	(0.0119)	(0.0206)
Pre-Round Output	0.0285	0.00236	0.0421*	0.00997	0.00115	0.0319**
	(0.0174)	(0.00773)	(0.0220)	(0.00864)	(0.00817)	(0.0127)
Constant	0.550**	0.221	0.623**	0.363	-0.331	1.504***
	(0.220)	(0.171)	(0.263)	(0.310)	(0.300)	(0.447)
Observations	672	243	429	336	121	215
Number of Subjects	96	79	94	48	37	46
Wald Chi^2	13.79	15.45	47.20	44.40	-	262.61

Robust standard errors in parentheses (clustered on group_id)

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

Random effects regression (period dummies included), reference category: Grade=1$_t$

Table 2.15: The impact of ratings on timeouts

		Game A				Game B			
		Pair I		Pair II		Pair I		Pair II	
		Payoffs (in €) for Player				Payoffs (in €) for Player			
	#	1	2	1	2	1	2	1	2
s	1	1.00	1.00	0.05	4.95	5.00	0.00	0.00	0.00
w	2	1.00	1.00	0.71	4.39	5.00	0.00	0.25	0.25
i	3	1.00	1.00	1.11	3.89	5.00	0.00	0.50	0.50
t	4	1.00	1.00	1.36	3.64	5.00	0.00	0.75	0.75
c	5	1.00	1.00	1.42	3.58	5.00	0.00	1.00	1.00
h	6	1.00	1.00	1.66	3.34	5.00	0.00	1.25	1.25
i	7	1.00	1.00	1.76	3.24	5.00	0.00	1.50.	1.50.
n	8	1.00	1.00	1.84	3.16	5.00	0.00	1.75	1.75
g	9	1.00	1.00	1.90	3.10	5.00	0.00	2.00	2.00
	10	1.00	1.00	1.93	3.07	5.00	0.00	2.25	2.25
p	11	1.00	1.00	1.96	3.04	5.00	0.00	2.50	2.50
o	12	1.00	1.00	2.03	2.97	5.00	0.00	2.75	2.75
i	13	1.00	1.00	2.07	2.93	5.00	0.00	3.00	3.00
n	14	1.00	1.00	2.09	2.91	5.00	0.00	3.25	3.25
t	15	1.00	1.00	2.12	2.88	5.00	0.00	3.50	3.50
	16	1.00	1.00	2.14	2.86	5.00	0.00	3.75	3.75
I	17	1.00	1.00	2.16	2.84	5.00	0.00	4.00	4.00
to	18	1.00	1.00	2.18	2.82	5.00	0.00	4.25	4.25
II	19	1.00	1.00	2.19	2.81	5.00	0.00	4.50	4.50
	20	1.00	1.00	2.21	2.79	5.00	0.00	4.75	4.75
	21	1.00	1.00	2.22	2.78	5.00	0.00	5.00	5.00
	22	1.00	1.00	2.50	2.50	5.00	0.00	5.25	5.25

Table 2.16: Eliciting social preferences - "#" indicates the unique switching point from pair I to pair II.

Dependent Variable:	Group Output					
	BaseBase vs FdsFds		BaseCost vs. FdsCost		BaseSabo vs. FdsSabo	
	Periods 1-8	Periods 9-16	Periods 1-8	Periods 9-16	Periods 1-8	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)
Pre-Round 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
Wald Chi^2	653.88	179.43	108.59	82.27	37.53	57.82

Robust standard errors in parentheses, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, in (3) and (6) clustered on group_id
Random effects regression on periodic group output

Table 2.17: The Impact of forced distribution on productivity - treatment variations

Dependent Variable:	Group Sabotage$_{t+1}$	
	BaseSabo (periods 1-8)	BaseSabo (periods 9-16)
	(1)	(2)
Span of Grades=1_t	1.454*	0.254
	(0.850)	(0.766)
Span of Grades=2_t	1.367	2.103***
	(0.833)	(0.773)
Span of Grades=3 or 4_t	2.742	3.349**
	(2.024)	(1.569)
Group Output$_t$	-0.0949***	-0.146***
	(0.0297)	(0.0373)
SD of Output$_t$	-0.158	-0.235
	(0.176)	(0.194)
Pre-Round Group Output	0.0277	0.0574
	(0.0586)	(0.0468)
Constant	6.479**	8.862***
	(3.096)	(3.316)
Observations	128	112
Number of Groups	16	16
Wald Chi^2	-	-

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

Random effects regression (period dummies included)

Reference category: span of grades=0_t

Table 2.18: The impact of deliberate differentiation on sabotage activity

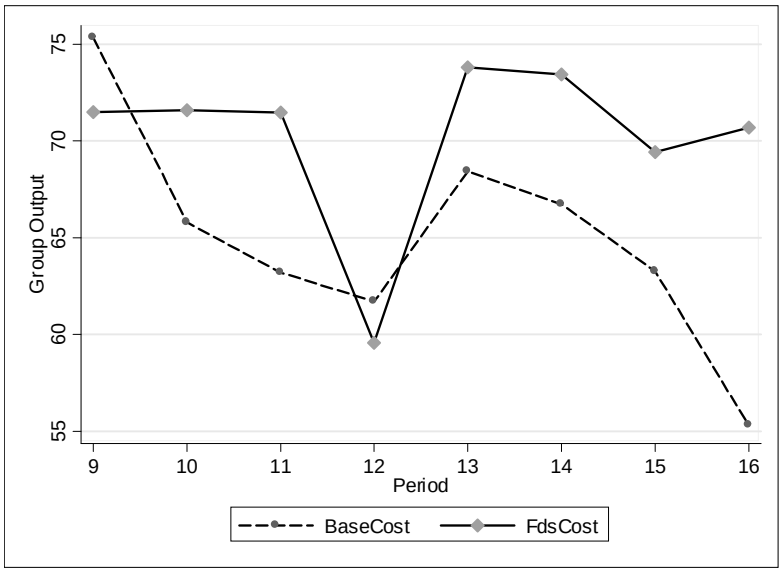

Figure 2.7: Group outputs over time across treatments when ratings are costly - in the last 8 periods

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 starts as soon as all participants 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 "timeout 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 timeout. So the worker loses 20 seconds per timeout since she/he cannot work on a block during this time. Please note that you cannot take a timeout 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/her 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:

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.

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

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.

Chapter 3

Heterogeneous Contestants and the Intensity of Tournaments - an Empirical Investigation[1]

3.1 Introduction

Tournaments where agents fight for a limited set of given prizes are omnipresent in day-to-day situations. One can for example observe promotion tournaments, competition for bonus pools (Baker et al. (1994), Rosen (1986), Rajan and Reichelstein (2006)) or tournaments concerning market shares and litigation contests between them (see for example Taylor (2003), Wärneryd (2000)). Also beauty contests, singing contests and sports competitions have the structure of tournaments (Amegashie (2009), Szymanski (2003)).[2]

As Lazear and Rosen (1981) have shown in their seminal article, rank-order tournaments can -under certain conditions- be the optimal design to induce first best effort levels if only ordinal information is available at reasonable costs. However, theory predicts that incentives are lower in heterogeneous tournaments, i.e. when contestants considerably differ with respect to ability or skill. In heterogeneous tournaments the underdog will shy away

[1] This chapter is based upon Berger and Nieken (2010).

[2] For an overview about tournaments and contests see for example Konrad (2009).

from competition as his chances of winning are comparably low. The opponent will anticipate this reduction of costly effort and decide to hold back effort as well. As a result, overall performance and, hence, the intensity of the tournament decreases. This effect is called the contamination hypothesis (e.g. Bach et al. (2009)). Since in practice contestants are seldom completely homogeneous, this prediction calls the frequent use and effectiveness of tournament schemes in firms and organizations into question. While the logic and effects of heterogeneous tournaments have been studied intensely in the theoretical literature (see among others Kräkel and Sliwka (2004)), only recently a growing body of papers test the theoretical predictions with non-experimental field data from sports contests (for instance Frick et al. (2008), Bach et al. (2009), for experimental evidence see Schotter and Weigelt (1992) or Harbring et al. (2007)).

The contribution of this paper to the existing literature is twofold: First, we analyze the impact of heterogeneity on the incentive effects of tournaments, using data from the TOYOTA Handball-Bundesliga.[3] We are the first to test the contamination hypothesis with data from handball, a game that provides measures necessary to test this particular prediction. We have collected data of two seasons, containing information on goals and fouls as well as ranks and odds from sports betting. Betting odds provide an excellent measure of the team's current ability as they contain all available information such as standings, recent performances, player injuries or transfers right before each game. They allow us to derive ex-ante winning probabilities which we then use to determine the heterogeneity of the match up. Furthermore, we use the number of 2-minute suspensions to approximate the intensity of the game. Our results confirm the contamination hypothesis and show that tournaments between heterogeneous contestants are significantly less intense. The results are robust to different measures of heterogeneity and sub sample analyses of the data. Second, we show that the overall decrease in game intensity is almost entirely driven by the reaction of the favorite team, i.e.

[3]Note that we, as well as Frick et al. (2008), consider the team as a unit and therefore rely on two-players models such as Lazear and Rosen (1981) instead of collective tournament models.

the favorite plays significantly less intense in asymmetric games while the underdog does not cease to exert effort "against all odds". In addition, we test if our proxy for game intensity is a suitable measure for our analysis. In line with the intuition that teams who put forth extra effort on the defensive end should be more likely to win, we find that the number of 2-minute suspensions is positively linked to the winning probability of the corresponding team.

3.2 Related Literature

Since objective measures for workers' abilities as well as effort or performance differences are rarely available, non-experimental field evidence on the contamination hypothesis is quite scarce. Studying professional sports data may help to fill this gap as sports contests often resemble very standardized tournament settings between two parties of which ability and performance proxies may be derived from game statistics. However, the studies which tested the contamination hypothesis with sports data do not provide unambiguous evidence in favor of it. Among the first studies, Ehrenberg and Bognanno (1990) analyze PGA golf tournaments and cannot clearly confirm the contamination hypothesis. They show that the stronger the opponent, the weaker the performance of a player. While this is in line with theory for participants performing below average, it violates theory for participants performing above average as they should be motivated by a higher quality opponent. Brown (2011) also uses data from PGA golf tournaments from 1999-2006 and shows that effort declines if a superstar (Tiger Woods) participates in the tournament. However, her findings are only significant for higher-skilled players but not for lower-skilled ones. Horse race studies like Lynch (2005) support the contamination hypothesis as does Sunde (2009) using tennis data. He also conducts a separate analysis for favorites and underdogs and finds that only underdogs are sensitive towards heterogeneity and reduce effort. In contrast to our paper, all these papers study individual sports. Bach et al. (2009) analyze data from the Olympic Rowing Regatta 2000 for teams and single skulls. They report higher effort levels in homogeneous groups, but also find

that only the favorites and not the underdogs react to heterogeneity. Bach et al. attribute this finding to the Olympic spirit which might motivate underdogs to do their best, irrespective their chances of success. Closest to our paper is the work of Frick et al. (2008) and Nieken and Stegh (2010). Frick et al. use data from the German soccer league. Employing betting odds to measure heterogeneity and red and yellow cards as proxies for effort, their main finding is in line with our results.

In this paper we go one step further and take the dynamic structure of tournaments into account by analyzing the teams' intensity of play separately for each half of the game. Our results show that ex-ante ability differences not only determine the intensity of a match at the beginning of the tournament but also towards the end, irrespective the halftime score. Nieken and Stegh analyze the effects of heterogeneity in the German Hockey League. Here, the number of minor suspensions also declines if contestants differ in their abilities. In contrast to our findings, they cannot confirm the contamination hypothesis for each third of the game separately. While we provide evidence that 2-minute suspensions may serve as a proxy for game intensity in handball, the previous mentioned work neglect this proof for their data.

The remainder of the paper is structured as follows. The next section describes the data set and our key variables. In section 3.4 we present our results and discuss our findings. Section 3.5 concludes the paper.

3.3 The Data

In our study we use professional sports data from the first "TOYOTA Handball-Bundesliga", the major handball league in Germany.[4] Our data set comprises all 612 league games from the seasons 2006/2007 and 2007/2008. For each game and each halftime we collected detailed information on the goals scored and penalties committed by both teams. We also gathered statistics on the number of spectators, size of venues and the two referees in charge of the game. Even though handball has become the second most popular sport

[4]The data are made publicly available and are downloadable in pdf-format under https://www.toyota-handball-bundesliga.de

in Germany[5], handball is still rather unknown outside European boarders. For the ease of comprehension, the next section briefly addresses the most important rules of the game.

3.3.1 The Game of Handball

In handball[6] two teams, each consisting of one goalkeeper and six field players, compete for two 30 minutes halves. By bouncing, passing and ultimately throwing a small ball into the goal of the opposing team, the team outscoring the opponent wins. In each season all 18 teams play every other team twice, once at home and once away. This amounts to a total of 34 league games for each team in each season. For each game, the winning team earns two championship points while the defeated team receives none. In case of a tie the two points are split up equally. The championship points determine the final league standing at the end of the season while the team with the most points wins the national title. In principle, all 9 top ranked teams may qualify for a European contest in the upcoming season[7] and up to three teams may lose their spot in the first national league. Since almost all final ranks have thus direct implications for the financial future of the ball club, incentives to win additional games are given throughout the entire season.

3.3.2 Heterogeneity Measures

The key independent variable needed to test our main hypothesis concerns the heterogeneity of the two agents (teams) competing in the tournament.[8] Intuitively, differences in team abilities should be reflected by differences in

[5] Among 1046 Germans, 40.7% respondents named handball the second most popular sport after soccer, followed by track and field and tennis with roughly 25% and 20% (Statista.de 2009).

[6] Handball is also known as team handball, Olympic handball or European handball. Note that American handball is a completely different game.

[7] This is the case when German teams have won all three European titles in the previous season as it happened in 2006/2007.

[8] We consider each game as a separate contest. As argued above, we believe that each game is important in itself. In our analysis we, however, try to control for seasonal trends and do separate regressions for different sections of the season.

current league standing. This measure may, however, yield noisy estimates early in the season when rankings usually fluctuate by a lot, not reflecting true abilities. Taking the difference in final rankings instead, one would assume constant ability differences over the course of season and ignore potential ups and downs caused by injuries or player transfers during the season. A more efficient indicator for ability differences between two teams can be derived from sport betting odds (see Fama (1970), Camerer (1989), Woodland and Woodland (1994), Levitt (2004) or Forrest et al. (2005) for a discussion about market efficiency in betting markets). Betting odds should be able to capture within-seasonal fluctuations of team ability more accurately than rankings. As Frick et al. (2008) and Deutscher et al. (2009), we use betting odds to proxy heterogeneity. Following their approach, we calculate the implicit winning probabilities of the respective teams based on betting odds from betexplorer.de. Taking the absolute difference of these probabilities results is our preferred measure of the match up's heterogeneity: "Het_Odds". This measure can take on any value between 0 (very homogeneous) and 1 (very heterogeneous contestants). The average in the sample corresponds to 0.49.[9]

3.3.3 Effort Measures

The other key variable needed to test the contamination hypotheses in our setting is team effort. The fact that the effort choice of the observational unit in a tournament is usually not directly observable poses a major empirical problem for testing the incentive effects of tournaments. In contrast to most firm data sets, sports data usually offer a larger amount of statistics. However, it is not always straightforward to decide upon which best reflect individual or team effort. Frick et al. (2008), for instance, argue that team effort in soccer is hard to measure with statistics kept on the offensive end of the game. The number of scored goals during a soccer match may not serve as a good proxy for team effort as scoring may simply result from a

[9]For a more detailed description please see the appendix or Frick et al. (2008) and Deutscher et al. (2009))

lack of defensive effort by the opposing team. The same argument holds for the game of handball. Similar to Frick et al. (2008), we believe overall team effort is - in our case - more accurately approximated by the effort put forth in defense which may be best captured by foul statistics.

Unlike in soccer, a foul in handball is not automatically considered unfair. In general, handball is considered a very physical game. Defensive players are allowed to stop the opponent by using body contact when they are in between the attacking player and their own goal. Even though the play is then interrupted and the offensive team regains possession of the ball, such a "fault" is considered a good defensive effort and is not penalized. In fact, if the defensive team can prevent the offense with "faults" from scoring for a long enough time, the referee may eventually call "passive play" urging the offensive team to wrap up its offensive effort. In this case, the defense is likely to prevent a goal and to get a chance to score themselves on the next possession. Harsher defensive attacks are, however, usually sanctioned by 2-minute suspensions. The player who committed the foul is then temporarily suspended from the game and leaves his team playing a man down for the next 120 seconds.[10] 2-minute suspensions are considered part of the game as they occur roughly 8 times during an average league game. They are thus more frequently ruled than yellow cards in soccer and should therefore be less prone to measurement errors such as poor referee judgments.

In our analysis 2-minute suspensions will serve as our proxy for team effort or the intensity of play. The idea behind this is as follows: A team who tries particularly hard to prevent the offensive team from scoring will play very physical defense. Often this additional effort on the defensive end will successfully prevent goals without players being sent off the court by the referees. However, sometimes these defensive attacks will be just outside the tolerated norm and result in a 2-minute suspension. Teams that lack defensive effort do not defend aggressively and are thus generally less likely to commit penalties.

One could also think of 2-minute suspensions as a proxy of destructive

[10]Each player may only receive two 2-minute suspensions. For his third 2-minute suspensions, he automatically receives a red card and is suspended for the rest of the game.

sabotage activity rather than effort. Nevertheless, subtle sabotage activities that successfully prevent goals and remain undetected in the majority of cases could also be considered good defensive effort. Fouling itself, without increasing "good" defensive efforts, is unlikely to be a rationale strategy as sabotage activities are likely to be detected by the referees. A team who decides to play illegal defense without increasing defensive efforts will constantly loose players due to 2-minute suspensions and thereby give up chances to win. Similar, if penalties were the result of frustration or a lack of good defensive effort, teams with more suspensions would be more likely to lose. In contrast to this, chapter 3.3 shows that more suspensions are associated with a higher likelihood to win. We therefore believe that 2-minute suspensions are more likely to be a "by-product" of high defensive effort rather than just an indicator of sabotage.

Since we cannot perfectly rule out that 2-minute suspensions also capture tendencies to sabotage the other team, we interpret total 2-minute suspensions per game as an indicator for the "intensity of the game" rather than "joint team efforts".[11]

Table 3.5 in the appendix provides summary statistics on the committed penalties as well as the main independent variables included in our upcoming analysis.

3.4 Results

In this section we present our main results. At first, we test if the intensity of the game is indeed predicted by the heterogeneity of the particular match up. In section 3.4.2 we report separate analysis on how ex-ante favorites and ex-ante underdogs react to ability differences in tournaments. Section 3.4.3 validates our measure of play intensity by explaining the outcome of the game by the number of 2-minute suspensions ruled against each team.

[11] Note that according to tournament theory, not only efforts but also sabotage activity should decrease in the heterogeneity of the tournament. Thus, even if penalties are a proxy for sabotage rather than effort or a mixture of both, theory would still predict less penalties in asymmetric contests.

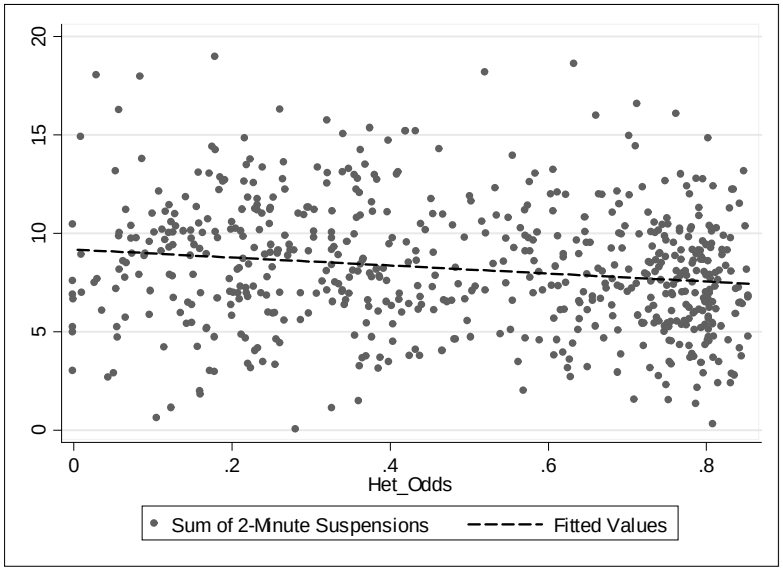

Figure 3.1: The relation between heterogeneity (based on betting odds) and the number of 2-minute suspensions

3.4.1 The Impact of Heterogeneity on the Intensity of the Game

A first descriptive picture on the relation between penalties and ability differences is given in figure 3.1.[12] The negative slope of the fitted value line is in support of the contamination hypothesis and reveals that the number of penalties indeed decreases in the heterogeneity of the match up. Of course, this conclusion may be far-fetched as it is only based on correlations without any further controls.

To investigate this relation in more detail, we apply regression analysis. As our dependent variable, i.e. the sum of 2-minute suspensions, is a count variable, we use Poisson regressions throughout our analysis.[13] Our main

[12]Note that as the data is count data, we used the Stata option "jitter" to make data points visible that would lie on top of each other otherwise.

[13]Figure 3.3 in the appendix shows that our dependent variable follows a poisson distribution. As shown in table 3.5, the variance of our dependent variable is only slightly larger than its mean, indicating that overdispersion is not a problem in our estimations. However, our results are also robust to other count model specifications such as negative binomial regressions as well as simple OLS regressions.

independent variable is the heterogeneity of the two contestants which is approximated by the absolute difference in winning probabilities (Het_Odds). For robustness checks, we alternatively use the absolute difference in final (Het_Final Rank) or current league standings (Het_Current Rank) as proxies for heterogeneity. Besides differences in team abilities, we control for several other factors that are also likely to affect the intensity of a game: As in any other team sport, certain match ups are more important for teams and fans than others. Such games usually take place between two local rivals and are referred to as "derbies". Since these games might in general be fought more intensely, we include a dummy variable (Derby) taking on the value 1 if a game can be classified as a derby and 0 otherwise.[14] Second, as pointed out in previous studies, the atmosphere created by fans could affect the players' actions on the court (see for instance Dohmen (2008)). Therefore, we additionally control for the absolute number of spectators attending the game as well as the percentage of taken seats. Given that some handball venues are much smaller than others, the latter variable gives us a better estimate on how relative attendance, e.g. if the venue is sold out, affects the intensity of the game. As certain teams might on average be more likely to commit fouls than other, dummy variables for both competing teams are included. To account for the course of the season, a dummy variable indicating the last 18 games of the season, a dummy indicating season 2007/2008 and a dummy of the interaction of the two (the last 18 games in the season 2007/2008) are added.[15] Finally, we also control for referee fixed effects in our estimations.

Table 3.1 displays our main results. Irrespective of the heterogeneity measure applied, we have highly significant evidence that the intensity of the contest - approximated by the sum of 2-minute suspensions per game - decreases in the heterogeneity of the match up. Holding all other variables constant, a one standard deviation higher absolute difference in winning probabilities of roughly 26%, is associated with a 7.6% decrease in the expected sum of

[14] We define a game as a derby if the cities of the two opposing teams are within 150 kilometer distance.

[15] We also ran regressions in which we included dummy variables for each day a match took place. Since it did not change our main results, we decided not to include these additional 60 dummies.

Dependent Variable:	Sum of 2-Minute Suspensions		
	(1)	(2)	(3)
Het_Odds	-0.3078***		
	(0.068)		
Het_Final Rank		-0.0233***	
		(0.004)	
Het_Current Rank			-0.0141***
			(0.004)
Derby	0.0854**	0.0739*	0.0915**
	(0.041)	(0.042)	(0.042)
Taken seats in %	0.1164	0.1073	0.0920
	(0.098)	(0.098)	(0.101)
Spectators/1000	0.0047	0.0032	0.0086
	(0.009)	(0.009)	(0.009)
Constant	2.0585***	2.1583***	2.0107***
	(0.124)	(0.128)	(0.130)
Observations	612	612	594
Pseudo R^2	0.08	0.08	0.08
Log pseudolikelihood	-1448.99	-1445.36	-1411.50

Poisson estimations, robust standard errors in parentheses
*** p < 0.01, ** p < 0.05, * p < 0.1, Further controls: referee dummies, home and away team dummies, season 2007/2008 (0/1), last 18 games of season (0/1), last 18 games in season 2007/2008 (0/1)

Table 3.1: The effect of heterogeneity on 2-minute suspensions

2-minute suspensions.[16] Similar, a one standard deviation larger difference in final standings (roughly 4 ranks) decreases the expected count of penalties by 9.2%. Moreover, penalties are more often ruled in games between two local rivals.

In table 3.6 in the appendix we opt for a nonparametric functional form of our main independent variable to allow for non-linearity of the effect. Here, we regress the dependent variable on the 2^{nd} to 5^{th} quintiles of our heterogeneity measures with the lowest quintile of heterogeneity being the reference category in all three specifications. The results show that the number of penalties constantly decreases in the degree of the heterogeneity of the match up. While column 2 and 3 suggest a rather linear relation between league standings and performance, the decrease in performance is somewhat convex when considering winning probabilities. Moving from the 1^{st} to the 2^{nd} quintile of winning probability differences, game intensity is only slightly and insignificantly smaller. However, the difference between the 1^{st} and 5^{th} quintile is highly significant and much larger than the significant difference between 1^{st} and 4^{th} quintile. Observing a game in the highest quintile of our heterogeneity measure Het_Odds (which on average corresponds to an 80% difference in winning probabilities) as opposed to a game in the lowest quintile of heterogeneity (which on average corresponds to a 13% difference in winning probabilities) decreases the expected count of suspensions by roughly 26%. A game in the 4^{th} quintile, as opposed to one in the 1^{st}, still decreases expected suspensions by 15%.[17]

In table 3.7 in the appendix we analyze the impact of heterogeneity on game intensity separately for each half of the game. One could argue that ex-ante ability differences become less important over the course of the game, as the halftime score provides both teams with a meaningful update of their current ability differences and the respective winning probabilities. We find that the number of suspensions significantly decreases in ability differences not only in the first but also in the second half. The effect of ex-ante ability

[16] To compute the percentage change in the expected count of our dependent variable, we use Stata's listcoef-package written by J. Scott Long and Jeremy Freese.

[17] The differences between the coefficients of the 5^{th} and the 4^{th} as well as the 4^{th} and the 3^{rd} quintile are significant.

differences is indeed somewhat smaller in the second 30 minutes. While a standard deviation increase in heterogeneity decreases the expected count of suspensions by 8.6% in the first half, the effect decreases to 6.2% in the second half.[18] The insignificant coefficient of "Halftime Score" further indicates that additional information on winning probabilities introduced by performance differences in the first half does not seem to affect game intensity in the second half of the tournament.[19]

One may argue that including all games of the season in the analysis is inappropriate as incentives to win could differ with respect to the progress made during the season.[20] Since we have considered each game as a separate contest, we do not fully account for the fact that each game is also embedded in a bigger contest, i.e. the championship race. Even though we argued that teams have considerable incentives to win games irrespective their current rank, we try to account for this simplification in our analysis by separately analyzing games in the first and in the second half of the season. If games toward the end of the season were perceived more or less important, the influence of heterogeneity should also vary across both sub samples. However, the coefficient of our main variable Het_Odds remains virtually identical and significant in both sub samples, suggesting that our main result is not sensitive to the round of play.[21]

Overall, we believe our results provide rather strong evidence in favor of the contamination hypothesis as predicted by economic theory (Lazear and Rosen (1981)) and confirmed by similar recent empirical studies (e.g. Frick et al. (2008), Bach et al. (2009) or Nieken and Stegh (2010)).

[18] However, the difference between both coefficients is not significant.

[19] Note that "Het_Odds" and "Halftime Score" are highly correlated. However, even if we exclude Het_Odds from the estimation, the difference in goals at the half has only a marginal significant impact on the suspensions ruled in the second half. Also the interaction of the two variables is insignificant.

[20] One could think that heterogeneity has a smaller effect late in the season when rankings are more certain than in the beginning.

[21] Regression tables are available upon request.

3.4.2 The Impact of Heterogeneity on Favorites and Underdogs

According to theory (see for instance Kräkel and Sliwka (2004)), favorites and underdogs[22] should not react differently to the heterogeneity of the match up. In games with heterogeneous contestants, the underdog has only little chances to win and should therefore refrain from providing much effort. The favorite should anticipate this reduction and lower his effort as well. Similar predictions can be derived regarding the sabotage activities of favorites and underdogs. As experimental studies have shown, underdogs often exert higher effort levels than theoretically predicted while the behavior in symmetric settings is roughly in line with theory (see Bull et al. (1987), Schotter and Weigelt (1992)). While Weigelt et al. (1989) find no significant differences when comparing effort levels of favorites and underdogs in unfair tournaments, Harbring and Luenser (2008) report that efforts of weak players are significantly higher than in symmetric settings if the prize spread is high. In a real effort experiment of van Dijk et al. (2001) players with lower ability try to win the tournament against a high ability contestant even though they lose in most cases and could avoid the tournament by playing a piece rate scheme.

Regarding sports data, the results are somewhat mixed. While Sunde (2009) shows that underdogs react stronger to heterogeneity than favorites, Bach et al. (2009) and Nieken and Stegh (2010) find the opposite. In their studies the favorite lowers his effort in more heterogeneous contests but the effort of the underdog remains nearly unchanged. One may argue that in sports, the general norm suggests not to give up irrespective the size of the deficit. In team sports this norm might be even more prominent as players do not want to let their teammates and coaches down. From an individual player's perspective, giving up could also result in being put to the bench in the next game. In contrast, the favorite team may dare to lower effort without risking social sanctions associated with a loss. Indeed the ex-ante favorite teams end up winning 75% of the sampled games. We

[22] We define favorites and underdogs according to the betting odds for each game.

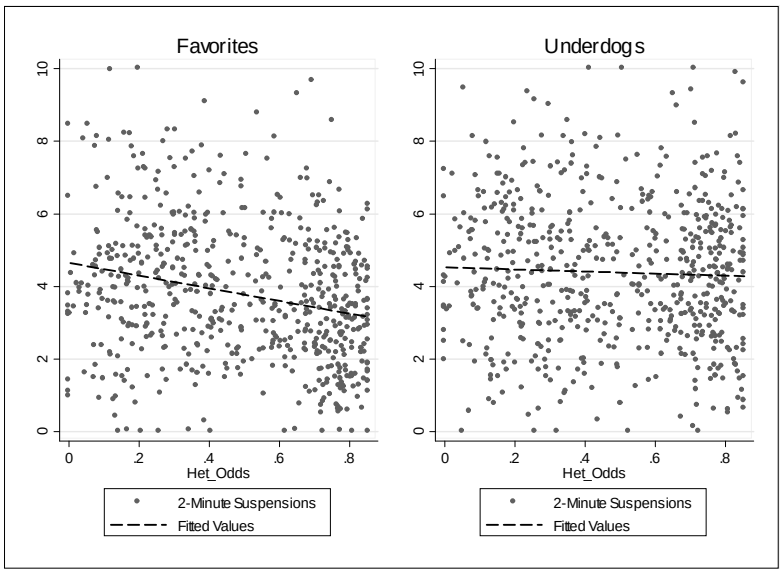

Figure 3.2: The relation between heterogeneity (based on betting odds) and 2-minute suspensions of favorites and underdogs

therefore expect favorites to be more willing to withhold effort (and sabotage) in heterogeneous contests than underdogs.

In figure 3.2 we show a scatter plot of committed 2-minute suspensions and the heterogeneity of the match up (Het_Odds) separately for favorites and underdogs. The picture seems to support the results found in Bach et al. (2009) and Nieken and Stegh (2010) as the favorites' number of penalties are substantially lower in heterogeneous contests while the right panel of figure 3.2 shows no systematic pattern for the underdogs. The overall decrease in the games' intensities, previously shown in figure 3.1, thus seems to be driven by the adjustments of the stronger contestants.

To confirm this impression, we run separate regressions for favorites and underdogs explaining the teams' committed penalties by the heterogeneity of the match up. Except for the dependent variable, the specifications in columns (1) and (3) are identical to our previous specification in table 3.2. In columns (2) and (4) we additionally test the linearity of the effect by regressing our dependent variable on the quintiles of our heterogeneity measure. The results indeed show that only the ex-ante favorite reacts to heterogeneity

by reducing the intensity of his play.

The coefficient in column (1) suggests that a one standard deviation increase in our measure "Het_Odds" reduces the expected count of 2-minute suspension of the favorite team by roughly 11%. Column (2) indicates that the favorite's reaction to the heterogeneity is monotone as indicated by the growing economic and statistical significance of higher quintile coefficients. However, the drop in effort is particularly pronounced in very heterogeneous games as the coefficient for the 5^{th} quintile is again nearly twice the size of the coefficient for the 4^{th} quintile. On average, the favorite's expected penalties are about 34% lower when the difference in the ex-ante winning probabilities falls into the 5^{th} quintile as opposed to the 1^{st} quintile. Interestingly, the coefficient for "Derby" is highly significant in both estimations, suggesting that favorites are willing to sacrifice additional effort when playing against one of their rivals. Columns (3) and (4) reveal that the underdog's play is hardly affected by ex-ante ability differences as all coefficients are economically and statistically insignificant. Table 3.8 in the appendix shows that this result is also reflected in the raw data. For the favorite, the average number of penalties decreases from 4.2 in the 1^{st} to 2.9 in the 5^{th} quintile of heterogeneity. For the underdog, the respective decrease ranges only from 4.6 to 4.3.

This finding is in line with Bach et al. (2009) and Nieken and Stegh (2010) but stands in sharp contrast to standard tournament theory. As mentioned above, this result may be attributed to social costs faced by inferior contestants for giving up. A similar argument is brought forward in a recent study by Fershtman and Gneezy (2011). In their field experiment the majority of students participating in running tournaments are unwilling to quit or drop out of the contest even when their prospects to win become negligible.

Some suggestive evidence for the existence of social sanctions imposed by the fans comes from table 3.9 in the appendix. Here we run separate regressions explaining the intensity put forth by the favorite during home and away games. The coefficient of "Het_Odds" in column (3) shows that the reduction of game intensity in heterogeneous matches seems larger when

Dependent Variable:	2-Minute Suspensions			
	Favorite		Underdog	
	(1)	(2)	(3)	(4)
Het_Odds	-0.4559***		0.0514	
	(0.106)		(0.098)	
2^{nd} Quintile		-0.0061		-0.0750
		(0.056)		(0.051)
3^{rd} Quintile		-0.1377**		0.0024
		(0.060)		(0.052)
4^{th} Quintile		-0.2159***		-0.0159
		(0.069)		(0.060)
5^{th} Quintile		-0.4154***		-0.0665
		(0.082)		(0.074)
Derby	0.1349**	0.1277**	0.0654	0.0651
	(0.059)	(0.059)	(0.050)	(0.049)
Seats taken in%	0.1003	0.0892	0.0871	0.0842
	(0.106)	(0.107)	(0.099)	(0.099)
Spectators/1000	-0.0166*	-0.0145	-0.0007	0.0022
	(0.009)	(0.009)	(0.008)	(0.008)
Constant	1.2104***	1.1716***	1.4272***	1.4506***
	(0.174)	(0.173)	(0.154)	(0.152)
Observations	612	612	612	612
Pseudo R^2	0.07	0.08	0.07	0.07
Log pseudolikelihood	-1153.01	-1149.55	-1189.80	-1188.74

Poisson estimations, robust standard errors in parentheses

*** p < 0.01, ** p < 0.05, * p < 0.1, Further controls: referee dummies, home and away team dummies, season 2007/2008 (0/1), last 18 games of season (0/1), last 18 games in season 2007/2008 (0/1)

Table 3.2: The effect of heterogeneity on 2-minute suspensions of favorites and underdogs

the favorite does not play in front of the home crowd. The coefficients of all quintiles of heterogeneity are larger and more significant at away games, but the differences across both sub samples are not quite significant when introducing interaction terms into a pooled estimation.

3.4.3 Testing our Measure of Game Intensity

How do we know that the number of 2-minute suspensions really serves as a good measure of game intensity? Increasing the intensity of play by putting forth more defensive effort and/or clever sabotage activities should, on average, increase a team's probability to win. If the number of suspensions is a result of these activities, more suspensions should be positively associated with the team's probability to win as well. If instead the number of suspensions reflects a lack of good defensive effort or the level of frustration, one would expect to see a negative relationship between penalties and winning probabilities as the team has to play a man down whenever a suspension is ruled.[23]

To validate our measure of game intensity, table 3.3 explains the outcome of the game by the share of penalties (0-100%) ruled against the ex-ante favorite team.[24] In specifications (1-4) our dependent variable is the difference in goals, i.e. the goals scored by the favorite team minus the goals scored by the underdog, while columns (5-8) explain the likelihood that the favorite team wins. If our line of thought is correct, an increase in the share of 2-minute suspensions should lead to a more favorable outcome for the corresponding team. This reasoning is partially confirmed in column (1) in which we explain the difference in scored goals using a simple OLS regression. Controlling for ex-ante winning probabilities and team fixed effects, the share of 2-minutes suspensions ruled against the favorite team has a positive and marginally significant impact on the difference in goals.

[23]However, results from soccer for instance indicate that even the permanent expulsion of a player does not necessarily lead to a disadvantage for the affected team (e.g. Caliendo and Radic (2006)).

[24]Note that the denominator of this measure already accounts for the overall intensity of the game as well as the number of fouls committed by the underdog.

In column (2) we again test for the linearity of this effect and see that the best outcome is achieved when the share of penalties rises to the 4^{th} quintile. In specifications (3) and (4) we repeat the previous estimations but restrict our sample to the 50% most homogeneous games. In these games, a team's marginal effort should have the largest impact on the outcome of the game. In the remaining games, ex-ante ability differences may be so large that the outcome of the game is hardly affected by effort or sabotage. Indeed, we find a much stronger and highly significant effect of the share of penalties on the difference in scored goals among homogeneous games. The linear estimate suggests that when the favorite's share of penalties increases by 20%, the difference in scored goals improves by 1.3 goals.

However, a team's effort or sabotage activities should be primarily directed toward winning the game and not toward outscoring the opponent by many goals. A more appropriate way to validate our measure is therefore to test its direct impact on the team's winning probability. Again, a simple descriptive statistic seems to suffice to support our argument. In the games which were won by the favorite, the average share of suspensions ruled against the favorite amounts to 46.4%, while in the games that were lost this number corresponds to 44.5%. In specification (5-8) of table 3.3 we further test this difference by regressing a dummy variable taking on the value 1 if the favorite team wins and 0 otherwise on the share of penalties and the control variables used in the previous specification. In specification (5-6) we again include all games in the analysis while (7-8) only include the most homogeneous games. The displayed coefficients are the marginal effects from a probit regression. Again the share of 2-minute suspensions ruled against the favorite significantly relates to the winning probability. The coefficient in column (6) implies that teams with a 10% higher share of 2-minute suspensions are 2.6% more likely to win. In homogeneous games this effect is almost 3 times as large.[25] The results in column (6) and (8) again imply that the effect of the share of 2-minute suspensions is more or less linear. The coefficient for the 5^{th} quintile, however, indicates that committing too many

[25] Among the 50% most homogeneous games, the favorite committed 51% of the penalties in the games he won and only 44% in the games which were lost.

penalties may eventually reverse this positive effect. Being too aggressive and thus committing too many fouls in relation to fair tackles will eventually harm the team.

Summing up, table 3.3 provides direct evidence that the number of 2-minute suspensions indeed reflect the intensity of play of a handball team which is reassuring for our reported main results. Note, however, that the interpretation of this result is unlikely to be that a team can increase its prospects to win by simply committing more fouls. Instead, teams who exert a lot of defensive effort are more likely to win but also more likely to commit fouls than teams who do not try hard to defend at all.

Dependent Variable:	Goals Favorite -Goals Underdog				1 if Favorite Wins			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All Games		Low Heterogeneity		All Games		Low Heterogeneity	
2- Minutes Share Favorite	0.0255*		0.0634***		0.0026**		0.0074***	
	(0.013)		(0.018)		(0.001)		(0.002)	
2^{nd} Quintile		0.9315		1.9889*		0.0407		0.1297
		(0.642)		(1.054)		(0.048)		(0.093)
3^{rd} Quintile		0.3208		1.7038*		0.0803*		0.2028**
		(0.576)		(0.914)		(0.045)		(0.085)
4^{th} Quintile		1.6631**		2.4201**		0.1208***		0.3179***
		(0.781)		(1.146)		(0.046)		(0.071)
5^{th} Quintile		1.1982*		2.7680***		0.0700		0.2081**
		(0.676)		(0.903)		(0.049)		(0.086)
Taken seats in %	0.1061	0.0750	1.0337	0.7863	0.1002	0.0866	0.1777	0.0422
	(1.158)	(1.175)	(1.895)	(1.912)	(0.118)	(0.119)	(0.220)	(0.213)
Spectators/1000	0.0725	0.0853	-0.2038	-0.1800	-0.0157*	-0.0149*	-0.0368**	-0.0282*
	(0.094)	(0.093)	(0.142)	(0.141)	(0.009)	(0.009)	(0.017)	(0.016)
Derby	-0.7632	-0.7589	-0.1018	-0.0758	-0.0360	-0.0488	0.0626	0.0599
	(0.640)	(0.643)	(0.944)	(0.952)	(0.062)	(0.064)	(0.110)	(0.104)
Het_Odds	8.5106***	8.3953***	6.6986***	6.7314***	0.5279***	0.5159***	0.7103***	0.6390**
	(1.161)	(1.174)	(2.334)	(2.365)	(0.098)	(0.098)	(0.269)	(0.262)
Constant	-5.1787*	-4.8539*	-7.7318**	-6.6033*	-	-	-	-
	(2.767)	(2.725)	(3.705)	(3.784)	-	-	-	-
Observations	611	611	305	305	611	611	304	305
R^2 or Pseudo R^2	0.34	0.35	0.23	0.23	0.21	0.21	0.17	0.16
Log pseudolikelihood	-	-	-	-	-274.95	-274.48	-169.08	-173.44

*** p < 0.01, ** p < 0.05, * p < 0.1, robust standard errors in parentheses, (1-4) OLS estimations, (5-8) Probit estimation (marginal effects reported), further controls: home and away team dummies, season 2007/2008 (0/1), last 18 games of season (0/1), season 2007/2008 in last 18 games (0/1)

Table 3.3: Game intensity as a determinant for game outcomes

3.5 Conclusion

Organizations often implement tournament schemes to induce incentives and decide about promotions of their employees. Indeed, tournaments can lead to first best effort levels but effort is predicted to decline if contestants are heterogeneous. Since in reality contests are seldom completely homogeneous, the effectiveness of tournaments in practice is called into question. As our analysis has shown, there is strong evidence in favor of the contamination hypothesis, i.e. heterogeneity between teams leads to a less intensive tournament. We find that especially the ex-ante favorite is likely to withhold effort while the underdog does not cease to exert effort "against all odds". In the game of handball or in team sports in general the latter result may be attributed to social or psychological costs the inferior contestant faces when not trying hard enough against an ex-ante dominant rival. However, in organizations such social costs may be absent or considerably lower as effort provision is not as publicly observable as it is in sports. In organizations underdogs might therefore also decide to spare costly effort when the prospects to win are considerably low.

To prevent this overall decrease in performance, firms should try to set up tournaments between contestants of similar ability. While in sports relegation systems or payroll caps help to ensure a competitive balance, firms can, for instance, match contestants with equal job profiles, educational background or tenure. If this is not possible, firms may consider handicapping the more able contestant (see for instance Lazear and Rosen (1981) or Knoeber and Thurman (1994)), adding absolute performance standards or refraining from using tournaments schemes at all.

3.6 Appendix

Heterogeneity Measure Calculation Example

To give an example of this calculation, consider the game between the TVB Lemgo and HSG Wetzlar which took place on December 12, 2007. Table 3.4 indicates that the home team TBV Lemgo was clearly favored by the bookmarkers. The corresponding odds imply that a bettor would receive 1.10 € for every Euro he or she placed on Lemgo. The unlikely case of a tie would yield 13.73 €, while a win of the away team would turn every Euro into 7.55 €. From the odds in table 3.4 it is straightforward to compute the payout ratio which can then be used to determine the winning probabilities of either team. The payoff ratio is given by the following equation:

$$\frac{1}{\frac{1}{\text{Odd Home Team wins}} + \frac{1}{\text{Odd Tie}} + \frac{1}{\text{Odd Away Team wins}}}$$

Example: 12/29/2007	Betting Odds	Probability
Win of TBV Lemgo	1.10	0.816
Tie	13.73	0.065
Win of HSG Wetzlar	7.55	0.119
Het_Odds	$\lvert 0.816 - 0.119 \rvert = 0.697$	

Table 3.4: Calculating the winning probabilities and deriving a heterogeneity measure from sports betting odds

In the given example the payoff ratio corresponds to 0.8974. Dividing this ratio by the payoffs connected to a win of either home or away team gives the winning probabilities of 0.82 and 0.12 respectively.

Variable	*Description*	*Mean*	*SD*	*Min*	*Max*
Round	Round of season	17.5	9.82	1	34
Referees	Id of referees	9.25	5.47	1	22
Goals	Sum of scored goals	59.18	7.49	38	87
Goals Favorite − Goals Underdog	Goals of the favorite − goals of the underdog	4.22	5.65	-15	27
Favorite Wins	Dummy=1 if favorite wins and 0 otherwise	0.75	0.44	0	1
Sum of 2-Minute Suspensions	2-minute suspensions per game	8.18	3.16	0	19
2-Minutes Favorite	2-minute suspensions of the favorite	3.79	1.89	0	10
2-Minutes Underdog	2-minute suspensions of the underdog	4.49	1.94	0	10
2- Minutes Share Favorite	2-minutes favorite / sum of 2-minute suspensions	45.88	15.91	0	100
Het_Odds	Abs. diff in winning probabilities	0.49	0.26	0.00	0.85
Het_Final Rank	Abs. diff in standings at the end of the season	6.33	4.11	1	17
Het_Current Rank	Abs. diff in current standings	6.09	3.96	1	17
Halftime Score (Diff)	Abs. diff in goals at halftime	3.76	2.83	0	13
Taken seats in %	Attendance relative to venue size	0.78	0.20	0.11	1
Spectators/1000	Spectators per game divided by 1000	4.76	3.11	1.00	19.40
Derby	Dummy=1 if cities of teams are within 150km distance	0.13	0.33	0	1

Table 3.5: Descriptive statistics of key variables

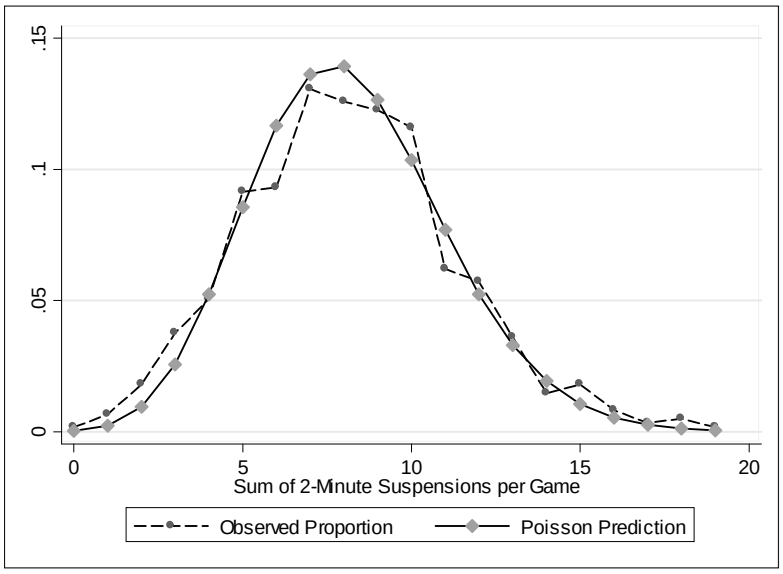

Figure 3.3: The distribution of 2-minute suspensions in the sample

Dependent Variable:	Sum of 2-Minute Suspensions		
	(1)	(2)	(3)
	Het_Odds	Het_Final Rank	Het_Current Rank
2^{nd} Quintile	-0.0511	-0.0586	0.0275
	(0.042)	(0.039)	(0.042)
3^{rd} Quintile	-0.0809*	-0.1451***	-0.0668
	(0.043)	(0.039)	(0.042)
4^{th} Quintile	-0.1585***	-0.1921***	-0.0998**
	(0.046)	(0.045)	(0.045)
5^{th} Quintile	-0.2960***	-0.2668***	-0.1550***
	(0.055)	(0.052)	(0.053)
Derby	0.0768*	0.0756*	0.0974**
	(0.041)	(0.042)	(0.042)
Taken seats in %	0.1471	0.1044	0.0945
	(0.097)	(0.097)	(0.100)
Spectators/1000	0.0016	0.0023	0.0085
	(0.009)	(0.009)	(0.009)
Constant	2.0319***	2.1151***	1.9667***
	(0.122)	(0.127)	(0.128)
Observations	612	612	594
Pseudo R^2	0.08	0.08	0.08
Log pseudolikelihood	-1444.93	-1444.00	-1409.77

Poisson estimations, robust standard errors in parentheses

*** p < 0.01, ** p < 0.05, * p < 0.1, Further controls: referee dummies, home and away team dummies, season 2007/2008 (0/1), last 18 games of season (0/1), last 18 games in season 2007/2008 (0/1)

Table 3.6: The effect of heterogeneity on 2-minute suspensions - different heterogeneity measures

Dependent Variable:	Sum of 2-Minute Suspensions			
	1^{st} Half		2^{nd} Half	
	(1)	(2)	(3)	(4)
Het_Odds	-0.3619***		-0.2467***	
	(0.100)		(0.088)	
Het_Final Rank		-0.0273***		-0.0192***
		(0.006)		(0.006)
Derby	0.0709	0.0544	0.0905*	0.0821
	(0.062)	(0.062)	(0.051)	(0.052)
Seats taken in %	0.0685	0.0574	0.1466	0.1396
	(0.175)	(0.176)	(0.134)	(0.135)
Spectators/1000	0.0025	0.0012	0.0069	0.0053
	(0.015)	(0.015)	(0.012)	(0.012)
Halftime Score (Diff)			-0.0075	-0.0061
			(0.006)	(0.006)
Constant	1.3152***	1.4350***	1.4275***	1.5069***
	(0.207)	(0.214)	(0.185)	(0.191)
Observations	611	611	611	611
Pseudo R^2	0.07	0.07	0.06	0.06
Log pseudolikelihood	-1147.77	-1145.74	-1290.33	-1288.78

Poisson estimations, robust standard errors in parentheses
*** p < 0.01, ** p < 0.05, * p < 0.1, Further controls: referee dummies, home and away team dummies, season 2007/2008 (0/1), last 18 games of season (0/1), last 18 games in season 2007/2008 (0/1)

Table 3.7: The effect of heterogeneity on 2-minute suspensions in each half

	2- Minute Suspensions (Game Averages)		
Het_Odds	Sum	Favorite	Underdog
1^{st} Quintile	8.73	4.18	4.55
2^{nd} Quintile	8.93	4.42	4.52
3^{rd} Quintile	8.11	3.84	4.28
4^{th} Quintile	7.94	3.64	4.30
5^{th} Quintile	7.16	2.88	4.29
All Games	8.18	3.79	4.39

Table 3.8: The relation between heterogeneity and 2-minute suspensions of favorites and underdogs

Dependent Variable:	2-Minute Suspensions Favorite			
	Favorite is Home Team		Favorite is Away Team	
	(1)	(2)	(3)	(4)
Het_Odds	-0.1796		-0.5322***	
	(0.216)		(0.197)	
2^{nd} Quintile		0.0243		-0.0014
		(0.075)		(0.102)
3^{rd} Quintile		-0.0914		-0.1364
		(0.095)		(0.113)
4^{th} Quintile		-0.1085		-0.2791**
		(0.116)		(0.133)
5^{th} Quintile		-0.2658**		-0.4082**
		(0.135)		(0.168)
Derby	0.2335***	0.2295***	0.0263	0.0123
	(0.073)	(0.073)	(0.087)	(0.092)
Seats taken in %	-0.0059	0.0306	-0.1398	-0.2015
	(0.168)	(0.165)	(0.284)	(0.282)
Spectators/1000	-0.0036	-0.0080	0.0098	0.0232
	(0.015)	(0.015)	(0.029)	(0.030)
Constant	1.2866***	1.2667***	1.3605***	1.2802***
	(0.190)	(0.188)	(0.294)	(0.300)
Observations	404	404	208	208
Pseudo R^2	0.09	0.10	0.09	0.09
Log pseudolikelihood	-747.01	-745.28	-383.03	-382.49

Poisson estimations, robust standard errors in parentheses
*** p < 0.01, ** p < 0.05, * p < 0.1, Further controls: referee dummies, home and away team dummies, season 2007/2008 (0/1), last 18 games of season (0/1), last 18 games in season 2007/2008 (0/1)

Table 3.9: The effect of heterogeneity on the favorites' number of 2-minute suspensions at away games

Chapter 4

Incentives and Cooperation in Firms - Field Evidence[1]

4.1 Introduction

Economic theory has often stressed that compensation based on team performance is accompanied by the danger of free-riding and consequently inefficient employee efforts. This problem has been discussed comprehensively in the theoretical and empirical literature.[2] However, several arguments in favor of team-based compensation were brought forward. A key argument is that under team-based incentive schemes employees should be more inclined to support teammates fulfilling their tasks which in turn is beneficial for the employer. Itoh (1991) and Itoh (1992), for instance, analyze formal models, showing that it can be worthwhile to base agents' rewards not only on individual but also on coworker performance when there is scope for mutual helping efforts.[3] In contrast, incentive schemes purely based on individual performance may reduce the willingness to help each other when

[1] This chapter is based upon Berger et al. (2011).

[2] See for instance Holmström (1982), Alchian and Demsetz (1972) and Newhouse (1973), or Prendergast (1999) for a survey.

[3] See also Holmström and Milgrom (1991), Drago and Garvey (1998) and Dur and Sol (2010). Within a dynamic framework Auriol et al. (2002) point out that team contracts also reduce potential negative effects of career concerns by weakening incentives to reduce colleagues‘ performance.

helping takes away time and resources from working on individual tasks (see for instance Lazear (1989), Drago and Garvey (1998), Encinosa et al. (2007), Burks et al. (2009)).

In this paper we investigate the connection between the structure of compensation schemes and the inclination to help coworkers empirically. We use a unique and representative employer-employee matched survey which was conducted by the Great-Place-to-Work Institute, a company specialized in conducting employee surveys, on behalf of the German Federal Ministry of Labor and Social Affairs in 2006. The data set is a sample of 305 German firms, containing company-level information about workers' and managers' performance-related payment schemes. In addition, in each firm an employee-survey has been conducted, containing detailed information about work satisfaction of approximately 36,000 workers.

We find that the intensity of team-based compensation schemes is significantly positively related to several measures of cooperation. However, neither incentives based on individual nor on firm performance affect cooperation among employees. The positive link between team-based incentives and cooperation is substantial: For example, a 10 percentage point increase in the share of team-based compensation (as a percentage of total compensation) is associated with an 11% increase in the number of employees who agree to the statement that in the firm *"you can count on people to cooperate"*. This relationship depends on workforce size and is stronger in smaller companies.

The data set also provides a direct survey question on the employees' general preference for helping others which allows us to disentangle selection from incentive effects. The effect remains basically unchanged when we control for helping preferences. Moreover, while there are strong inter-industry differences in the preference for helping, we find no differences between firms with and without team compensation schemes. Hence, we can rule out that the results are driven by the self-selection of more cooperative employees into organizations that use team-based incentives.

In addition, we investigate the connection between the structure of incentive schemes and absenteeism. In line with the previous observations, we also

find evidence for less absenteeism in the presence of team incentive plans.

While there is now some consistent field evidence showing positive effects of team incentive plans on performance (e.g. Jones and Kato (1995), Knez and Simester (2001), Hamilton et al. (2003), Bandiera et al. (2010b) Jones et al. (2010)), there are, to the best of our knowledge, only a very limited number of studies focusing on the link between team incentives and helping on the job. Drago and Garvey (1998) detect no relationship between helping efforts and the existence of piece rates or profit sharing using data from a survey of nonsupervisory employees at 23 Australian workplaces where helping effort is measured using responses to a survey question "To what extent do your fellow employees refuse to let others use their equipment, tools, or machinery?". Heywood et al. (2005) analyze the relationship between profit sharing and cooperation with the 1995 wave of the German Economic Panel and find a positive association between profit sharing and the perception that employees get along well with their colleagues. While these studies only use binary information, our data set contains information about the presence and the strength of individual, team- and firm-based performance pay which allows us to distinguish between the effects of these three components which typically make up incentive plans.

Our second result, that team incentives are associated with lower absenteeism rates, is in line with recent findings by Knez and Simester (2001), Bhattacherjee (2005) and Roman (2009). A possible explanation is given by Kandel and Lazear (1992) who identify team incentives as a determinant for peer pressure. While evidence from field studies (Ichino and Maggi (2000), Sacerdote (2001), Mas and Moretti (2009), Bandiera et al. (2010a)) or experiments (see for instance Falk and Ichino (2006), Mohnen et al. (2008)) highlight the importance of peer effects in general, field evidence on the connection between the structure of incentive schemes and peer effects is still rather scarce.

The remainder of the paper is organized as follows: In the next section we present the two data sets, the matching procedure and our hypotheses. Section 4.3 presents our main results. To meet endogeneity issues often raised in cross-sectional research designs, this section also includes several

subsample analyses and control specifications. In section 4.3.3 we present our findings concerning absenteeism and team incentives, before concluding in section 4.4.

4.2 Data and Hypotheses

Our data source is a 2006 employer-employee matched survey conducted by the Great-Place-to-Work Institute and the German Federal Ministry of Labor and Social Affairs. The data set is a representative sample of 305 German firms employing a minimum of 20 workers. In each firm, the management provided company-level information on organizational facts, corporate values as well as on various HR practices such as trainings, benefits and compensation. Most of this information is provided separately for managers and workers in each firm.[4]

In addition to this firm-level information, a representative employee-survey was conducted at each sampled firm, yielding over 36,000 observations in total. Among others, the employee survey includes 58 standardized items to be answered on a 5-point Likert scale which are designed to measure the level of trust, pride, and cooperation within firms. More precisely, the items focus on the relationship between employees and management, the work environment, and the relationship between employees. In our analysis we focus on the last aspect, i.e. the perceived level of cooperation among colleagues.

Due to the random sampling process, the 305 firms are almost evenly spread across the different industries in Germany. The majority of the sampled firms are small or medium sized. While the average number of employees amounts to 430, the median is at 157. However, roughly 10% of the firms employ more than 1,000 workers including the largest firm in the sample with 14,000 workers.

Previous studies (e.g. Drago and Garvey (1998), Heywood et al. (2005)) mainly relied on binary information about *whether* workers participate in firm profits. Our data set allows a more in depths analysis on *how much*

[4]More specifically, answers were provided for employees in supervisory function and for the largest group of non-managerial employees, i.e. the core occupational group.

employees benefit from economic outcomes and which pay components drive the effects. Each firm stated whether wages for managers and workers in the corresponding firm include a performance-related pay component. For both, managers and workers, we know the share of the average wage (in %) which is determined by performance-related pay (henceforth PRP). Furthermore, firms reported how much (in %) of total PRP is determined by either individual, team, or firm performance. Multiplying these numbers, we derive the fractions (in %) of the total wage that are based on the three different types of PRP.

Figure 4.1 gives a descriptive overview of PRP usage across industries, showing the share of firms using PRP. While the majority of sampled firms use variable pay components for managers, the use of worker PRP varies from only 6% of all organizations in the Public Sector to 71% in Financial Services. In total 109 out of 294[5] firms use PRP for their core occupational group. Figure 4.2 shows the composition of workers' incentives across industries. Though firm- and team-based variable compensation is quite common, individual incentive schemes have the prominent role. Roughly 55% of variable wage components are based on individual performance. Table 4.1 reports the average strength of incentives for the subset of firms who use at least one type of worker PRP. The mean magnitude of worker's incentive pay amounts to roughly 12% of the fixed wage. While workers' incentive pay is mainly based on individual performance, the largest fraction of managers' incentives is determined by the economic success of the company as a whole. For both groups, team incentives are relatively low. In firms using worker PRP, team incentives only account for 18% of total incentives and thus for only 2.2% of the total average wage.

Complementing the firm level information provided by management, we exploit the employee surveys conducted in each firm to measure the degree of cooperation among the workforce.[6] Table 4.2 shows 4 items of the employee survey which reflect workers' perception of teamwork and team atmosphere

[5] 11 out of the 305 sampled firms did not provide information on PRP.

[6] In firms with less than 500 employees all employees were asked to participate. In larger firms a representative 500-employee sample was drawn.

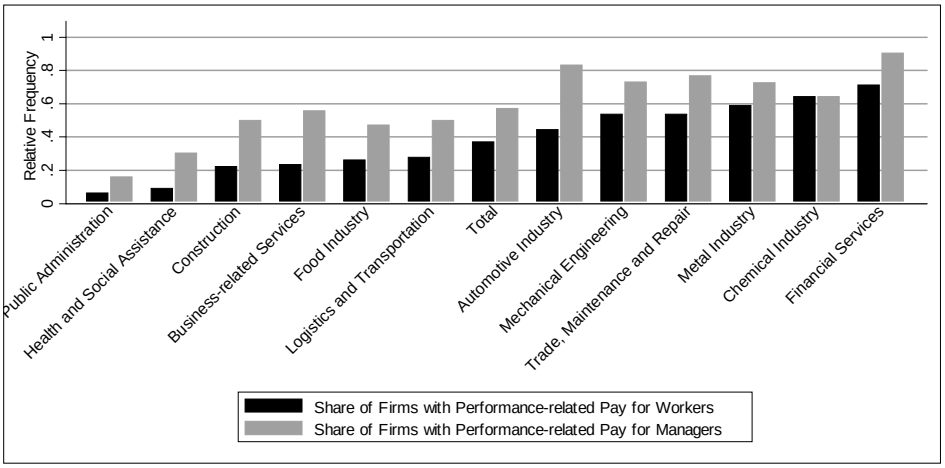

Figure 4.1: Utilization of performance-related pay across German industries

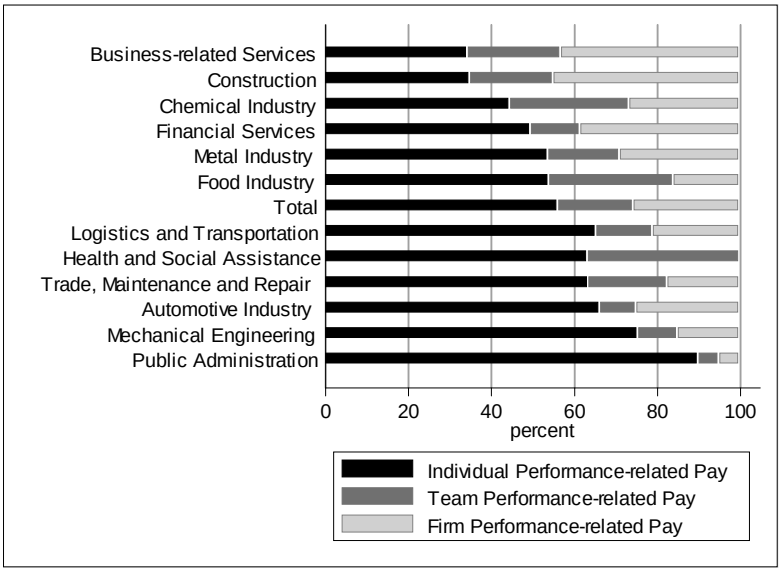

Figure 4.2: Structure of performance-related pay across German industries

Variable	Description	Workers	Managers
Individual PRP	% of Individual PRP on Average Wages	7.5%	4.8%
Team PRP	% of Team PRP on Average Wages	2.2%	4.4%
Firm PRP	% of Firm PRP on Average Wages	2.6%	8.4%
Total PRP	Total Percentage of PRP on Average Wages	12.4%	17.6%

Table 4.1: Utilization of performance-related pay in the sample

within a firm. All items use the same 5-point Likert scale ranging from 1 "*almost always untrue*" to 5 "*almost always true*" and refer to the company as a whole. The table displays simple descriptive statistics of responses given by full-time workers in all sampled firms.[7] The top-box column shows the percentage of workers who affirm a statement by choosing 4 or 5 on the 5-point scale. Overall, 54.6% of the responders affirm the statement "*You can count on people to cooperate*". The share of workers in a firm agreeing to an item serves as a dependent variable and is coded between 0 and 100.

Variable	Description	Top-Box	Sd
(1) Cooperate	"You can count on people to cooperate"	54.6%	17.6
(2) Care	"People care about each other"	52.5%	18.1
(3) Team Spirit	"There is a "family" or "team" feeling here"	45.9%	19.2
(4) Backstab	"People avoid politicking and backstabbing"	47.6%	18.1

Table 4.2: Survey items approximating cooperation

Detailed firm level information on PRP and suitable measures for team work in the firm allow for testing the relationship between incentives and the level of cooperation. We expect cooperation in firms to be positively affected by team incentives. The relation between individual incentives and cooperation is less clear cut. If supplying helping effort raises the costs for supplying 'private' effort, individual incentives reduce the inclination to help coworkers. If costs for helping effort are, however, independent of the costs of 'private' effort supply, individual incentives do not affect helping on the job (see Itoh (1991)). Incentives based on firm performance only gradually differ from team incentives, since a firm can been seen as a large team. However, the marginal effect on firm performance should be much smaller than the effect on team performance measures. Second, peer pressure is less likely to be sustainable as mutual monitoring becomes impracticable in larger teams. Hence, we expect to find a weaker relationship between firm level incentives and cooperation.

[7] Full-time employees with non-supervisory function are most likely to correspond to "the largest share of employees in the firm" addressed in the management survey questions. In the analysis of worker pay schemes on cooperation we therefore restrict our analysis to the answers given by this group.

Several other firm specific characteristics might also contribute to the level of perceived cooperation. As laid out above, the level of cooperation within a firm should be influenced by the number of workers constituting a team unit. We use the number of hierarchical levels to control for potential differences in team unit size across firms. For a given workforce size, more hierarchical levels should positively affect cooperation among workers due to a smaller average team size. In contrast, more hierarchical levels might also entail stronger promotion based incentives which in turn generate incentives to refrain from helping or even to sabotage colleagues (see Lazear (1989) and Drago and Garvey (1998)).

Moreover, the effect of team performance pay on cooperation might be mitigated by workforce size. Large firms tend to offer a greater variety of workplaces and development possibilities. Employees can avoid peer pressure by changing team, division, or location. Workers in small firms have fewer within-firm exit options and are exposed to potential peer pressure to a higher degree. Therefore, team-based compensation in small firms may lead to higher degrees of cooperation. In small firms employees are also more likely to interact in the future because the number of potential coworkers is limited. Hence, behavioral responses to team incentives do not only affect present but also future interaction with colleagues and should therefore foster cooperation. Che and Yoo (2001), for instance, show that under team incentives a higher frequency of future interactions increases productivity in a repeated game.

4.3 Results

4.3.1 Team Incentives and Cooperation

To study the relationship between incentives and the level of cooperation we match the firm-level information obtained in the management survey to the percentage of workers‘ affirmative answers to the survey items on cooperation. We then estimate the relation between incentive scheme structure and the percentage of workers agreeing to these cooperation items with OLS re-

gressions.[8] We control for firm characteristics such as firm size, industry and the presence of a works council. As noted above, we include the number of hierarchical levels to approximate team unit size when firm size is controlled for.

Table 4.3 presents our main results.[9] Team PRP is indeed significantly and positively related to all cooperation items. In economic terms, a 10% point higher team PRP is associated with a 6 percentage point increase in the fraction of affirmative answers to the item *"You can count on people to cooperate"*. The predicted fraction of employees agreeing to that statement at the mean of all other explanatory variables is equal to 54.3% when there is no team PRP. This fraction increases by about 11% to 60.3% of all employees when team PRP is 10% instead. The effect is of similar magnitude for all four items.

However, we do not find any relationship between our measures of cooperation and the strength of firm incentives. Also, higher individual incentives do not seem to be harmful for the perceived degree of cooperation. This indicates that there are no or rather low substitution effects between individual and helping efforts.[10] Furthermore, it is interesting to note that employees state a higher rate of cooperation if their firm organizes work in smaller team units as suggested by the positive coefficient of *Hierarchical Levels* at a given firm size.

According to our hypothesis, we should expect a stronger impact of team incentives in smaller firms. Table 4.4 captures the interaction between incentive pay and firm size. Note that for 3 of 4 items, the effect of team compensation negatively interacts with workforce size. The relation between team incentives and cooperation is thus particularly strong in small firms and tends to diminish with workforce size. In our linear interaction the relationship between team incentives and cooperation vanishes at a workforce

[8] Note that that there are nearly no observations of the dependent variables at the boundary of the $[0, 100]$ interval. Hence, tobit regressions lead to nearly identical results.

[9] Table 4.9 gives descriptive statistics of all explanatory variables.

[10] In the notation of Itoh's (1991) model, the employees seem to rather have "task specific" disutility of effort such that their individual cost functions are rather additively separable in costs for individual efforts and costs for helping efforts with vanishing cross derivatives.

All Firms

Dependent Variable:	Cooperate	Care	Team Spirit	Backstab
	(1)	(2)	(3)	(4)
Individual PRP	0.023	0.050	0.147	0.172
	(0.134)	(0.118)	(0.153)	(0.135)
Team PRP	0.599***	0.487***	0.620***	0.575***
	(0.227)	(0.182)	(0.167)	(0.195)
Firm PRP	0.056	0.170	0.380	-0.126
	(0.344)	(0.420)	(0.527)	(0.572)
Hierarchical Levels	3.268***	3.529***	1.790**	1.824*
	(0.827)	(0.962)	(0.877)	(1.089)
Works Council	-4.734**	-4.524*	-5.007*	-9.103***
	(2.329)	(2.561)	(2.808)	(2.459)
Constant	55.355***	51.677***	48.863***	53.493***
	(5.531)	(6.024)	(5.752)	(6.193)
Observations	281	281	281	281
R^2	0.22	0.19	0.15	0.18

*** p < 0.01, ** p < 0.05, * p < 0.1, robust standard errors in parentheses
OLS regression: further controls: 2 firm size dummies and 11 industry dummies
Reference category: 0-99 employee firm in the food industry

Table 4.3: Performance-related pay and cooperation among workers

size of approximately 400.

All Firms

Dependent Variable:	Cooperate	Care	Team Spirit	Backstab
	(1)	(2)	(3)	(4)
Individual PRP	0.032	0.058	0.106	0.211
	(0.177)	(0.153)	(0.200)	(0.181)
Team PRP	0.679**	0.667***	0.787***	0.774***
	(0.289)	(0.220)	(0.205)	(0.206)
Firm PRP	0.216	0.273	0.354	-0.218
	(0.458)	(0.553)	(0.686)	(0.757)
Workers/100	0.001	0.002*	0.001	0.004***
	(0.002)	(0.001)	(0.002)	(0.001)
Individual PRP × Workers/100	0.018	0.023	0.022	0.001
	(0.021)	(0.019)	(0.023)	(0.022)
Team PRP × Workers/100	-0.068	-0.176**	-0.190**	-0.196**
	(0.105)	(0.081)	(0.095)	(0.083)
Firm PRP × Workers/100	-0.041	-0.010	0.015	0.072
	(0.066)	(0.065)	(0.076)	(0.085)
Hierarchical Levels	3.243***	3.475***	1.739*	1.759
	(0.835)	(0.956)	(0.885)	(1.110)
Works Council	-4.894**	-4.780*	-5.139*	-9.403***
	(2.364)	(2.593)	(2.842)	(2.492)
Constant	55.462***	52.033***	49.297***	53.969***
	(5.619)	(6.090)	(5.826)	(6.304)
Observations	281	281	281	281
R^2	0.22	0.20	0.16	0.19

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

OLS regression: further controls: 11 industry dummies

Reference category: 0-99 employee firm in the food industry

Table 4.4: Performance-related pay, cooperation among workers and firmsize

In a further robustness check, we consider two more homogenous subsamples of firms. First, we restrict the analysis to firms which use at least one form of performance-based pay. In the next step, we consider only firms which use team incentives for their employees. The left panel of table 4.5 shows results for firms which use at least one type of PRP. We again find a positive and significant relationship between team incentives and cooperation, comparable in magnitude and statistical significance with the proceed-

ing analysis. The right panel displays a similar picture for the subsample of firms using team PRP. Even in this drastically reduced sample, our main result remains robust across all four items.[11]

[11]Due to the reduced sample of 40 firms, we do not include industry dummies in these specifications. In the preceeding analysis industries showed little statistically significance. Including industry dummies here yields a 20 regressor-40 observation regression with no significant estimates.

	Firms Using Worker PRP				Firms Using Team PRP			
Dependent Variable:	Cooperate	Care	Team Spirit	Backstab	Cooperate	Care	Team Spirit	Backstab
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Individual PRP	-0.0414	0.0686	0.133	0.215*	-0.252	0.034	0.098	-0.021
	(0.118)	(0.115)	(0.101)	(0.110)	(0.211)	(0.184)	(0.175)	(0.214)
Team PRP	0.535**	0.481**	0.531***	0.418*	0.540***	0.628***	0.801***	0.718***
	(0.239)	(0.204)	(0.191)	(0.239)	(0.191)	(0.179)	(0.140)	(0.235)
Firm PRP	-0.336	-0.195	-0.0324	-0.721	0.603	0.557	0.681	0.832
	(0.350)	(0.441)	(0.495)	(0.549)	(0.789)	(0.664)	(0.635)	(0.658)
Hierarchical Levels	3.419**	3.223**	2.002*	1.499	-2.278	-3.297	1.408	0.110
	(1.611)	(1.257)	(1.157)	(1.145)	(6.045)	(5.117)	(4.075)	(7.203)
Works Council	-8.980***	-9.543***	-8.242**	-11.60***	6.088**	0.729	2.168	-0.120
	(3.379)	(3.538)	(3.717)	(4.306)	(2.959)	(2.359)	(2.399)	(2.313)
Constant	49.65***	41.31***	38.94***	43.36***	45.324***	41.507***	29.013***	39.212***
	(4.384)	(5.736)	(5.307)	(5.438)	(8.341)	(7.965)	(6.325)	(7.307)
Observations	101	101	101	101	40	40	40	40
R^2	0.43	0.46	0.45	0.46	0.22	0.25	0.28	0.33

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

OLS regression: further controls: 2 firm size dummies and in specifications (1)-(4) 11 industry dummies

Reference category: 0-99 employee firm in the food industry

Table 4.5: Performance-related pay and cooperation among workers - subsample analysis

As the management survey contains detailed information on other management practices, we are able to control for further firm characteristics that are potentially confounding factors: The fraction of part-time employees, for instance, may affect the intensity of daily interaction of the workforce. Information about the wage level captures the company's wage policy and the attractiveness of a workplace. Whether a firm is currently downsizing or upsizing may have effects on the level of cooperation and may also affect the structure of compensation. Trainings could foster social interaction among the workforce and thereby affect cooperation. The presence of systematic female career support reflects the company's antidiscriminatory efforts and attempts to create a fair working environment. Furthermore, the general working climate, captured by the share of workers who are satisfied with their current job, may not only influence cooperative behavior but could also be influenced by the company's wage scheme. Table 4.10 shows estimates for column 1 of our basic specification from table 4.3 and the additional controls discussed above.[12] The effect of team PRP remains statistically and economically stable over all specifications, indicating a robust relationship between team PRP and cooperation among the workforce.[13]

4.3.2 Incentives or Self-Selection?

It is important to understand the key mechanism by which team incentives affect cooperation in more detail. Indeed, a given set of employees should have stronger incentives to cooperate if team performance is rewarded. But in addition, self-selection could also play a role as workers with preferences for cooperation may self-select into firms with team incentives. Then cooperation should increase simply due to the different composition of the workforce. Lazear (2000), for instance, showed in his seminal study on the effect of piece rates on productivity that about half of the productivity effect was due to self-selection. Moreover, recent laboratory studies (e.g. Cadsby et al. (2007), Dohmen and Falk (2011), Eriksson and Villeval (2008)) suggest that

[12] Regressions for all other items show almost identical patterns.

[13] The substantially reduced number of observations in the last column results from missing values in firms' training or gender career programs.

payment scheme design causes sorting effects, not only with respect to agents' abilities but also to their social preferences.

To investigate the self-selection argument in our data, we explore another subsection of the employee survey in which employees were asked which aspects of a job are important to them in general. Besides job security, high income or promotion opportunities, workers were also asked: "*How important is it for you to have a profession in which you can help others?*" which should capture an individual's general willingness to help others. If self-selection with respect to the specific structure of performance pay plays a role, we should expect the fraction of workers with a preference for helping to be higher in firms that tie rewards to team or firm performance. Including the fraction of workers with a preference for helping as an additional control in our baseline specification should then also reduce the coefficient of team PRP.

In the models reported in table 4.6, we first regress the share of workers in a firm stating that a job in which one can help others is important or very important to them on the structure of incentive pay and our set of standard firm controls. We again run the regressions for the entire sample but also for the Using-PRP and Using-Team-PRP subsamples. In none of the specifications neither individual, team nor firm PRP significantly explain the share of employees to whom helping is important.[14]

We also include this measure of the employees' general preference for helping in our basic OLS estimation to control for the share of cooperative workers in the firm. The results are displayed in table 4.7 and show that the coefficients of our variables of interest remain almost unchanged. Hence, we conclude that self-selection seems to be no key driver for the positive relation between team incentive schemes and cooperation in our data.

Interestingly, the distribution of cooperative preferences is quite heterogeneous across industries, as displayed in figure 4.3 where we graph the coefficients of the industry dummies included in table 4.6. Maybe not surprisingly, the share of cooperative workers is largest in health and social assistance and

[14]Note that we do find, for instance, that the share of workers stating that a high income is important to them increases in the strength of individual incentives.

Dependent Variable:	Preference for Helping		
	All Firms	Using PRP	Using Team PRP
	(1)	(2)	(3)
Individual PRP	0.0218	-0.0524	0.261
	(0.114)	(0.149)	(0.407)
Team PRP	-0.324	-0.184	-0.483
	(0.284)	(0.280)	(0.304)
Firm PRP	-0.177	-0.111	-0.546
	(0.327)	(0.450)	(0.955)
Hierarchical Levels	0.0146	-0.813	-2.756
	(0.519)	(1.015)	(5.109)
Works Council	-4.692***	-7.636**	-18.31**
	(1.698)	(3.706)	(6.991)
Constant	88.28***	82.15***	93.70***
	(2.549)	(4.849)	(11.97)
Observations	281	101	40
R^2	0.34	0.35	0.63

*** p < 0.01, ** p < 0.05, * p < 0.1, robust standard errors in parentheses
OLS regression: further controls: 2 firm size dummies
and 11 industry dummies
Reference category: 0-99 employee firm in the food industry

Table 4.6: Performance-related pay and self-selection

All Firms

Dependent Variable:	Cooperate	Care	Team Spirit	Backstab
	(1)	(2)	(3)	(4)
Individual PRP	0.022	0.052	0.146	0.174
	(0.138)	(0.115)	(0.154)	(0.133)
Team PRP	0.622***	0.464**	0.631***	0.555***
	(0.222)	(0.188)	(0.172)	(0.198)
Firm PRP	0.069	0.157	0.386	-0.137
	(0.336)	(0.432)	(0.521)	(0.579)
Hierarchical Levels	3.267***	3.530***	1.790**	1.825*
	(0.823)	(0.959)	(0.878)	(1.096)
Works Council	-4.408*	-4.853*	-4.853*	-9.394***
	(2.331)	(2.593)	(2.818)	(2.442)
Preference for Helping	0.069	-0.070	0.033	-0.062
	(0.111)	(0.092)	(0.109)	(0.101)
Constant	49.226***	57.860***	45.955***	58.971***
	(10.057)	(9.088)	(10.157)	(9.928)
Observations	281	281	281	281
R^2	0.22	0.19	0.15	0.18

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

OLS regression: further controls: 2 firm size dummies and 11 industry dummies

Reference category: 0-99 employee firm in the food industry

Table 4.7: Performance-related pay, self selection and cooperation among workers

lowest in financial and business-related services. Since our helping preference measure delivers plausible results for workers sorting into different industries, we are confident about our conclusion that incentive schemes do not lead to self-selection according to these preferences.

4.3.3 Team Incentives and Absenteeism

Having investigated the relationship between team incentives and perceived cooperation, we further test whether this positive relation is also reflected in more objective performance measures. A key figure that most management representatives (259 out of 305) were able to provide is the workers' average number of missed work days. In our sample, a worker missed on average 9 days of work.

Absenteeism is likely to decrease with rising individual incentives. Moreover, absenteeism is also predicted to decrease with higher team incentives. Recent studies have indicated that team incentives and increased peer pressure can effectively prevent workers from staying at home (Knez and Simester (2001), Bhattacherjee (2005) and Roman (2009)). Alternatively, if team incentives strengthen team spirit and cooperation, as suggested by our study, this mechanism might additionally reduce absenteeism. In a sense, a well functioning team may prevent workers from letting their colleagues down. To further test the economic importance of team incentives, we regress yearly absenteeism days on the incentive structure observed in each firm.

Table 4.8 shows that higher team incentives are indeed linked to fewer absent days. In our first specification, a 10% point increase in team PRP is associated with 1.4 fewer absent days per worker and year. Controlling for job satisfaction and average workforce age in specification 2, a 10% point higher team PRP comes along with one absence day less. Interestingly, individual PRP is far from statistical significance in both specifications. On the right side of table 4.8, we again restrict the analysis to firms who use PRP for their workers. Even in this substantially smaller sample, the main result that higher team incentives are associated with less absenteeism remains significant.

Dependent Variable:	Average Absent Days			
	All Firms		Firms Using PRP	
	(1)	(2)	(3)	(4)
Individual PRP	0.023	0.029	-0.044	-0.046
	(0.044)	(0.045)	(0.067)	(0.068)
Team PRP	-0.132**	-0.097*	-0.228**	-0.201*
	(0.059)	(0.059)	(0.107)	(0.113)
Firm PRP	0.010	-0.017	-0.174	-0.197
	(0.097)	(0.099)	(0.165)	(0.176)
Works Council	2.738***	2.117***	2.716*	2.325
	(0.746)	(0.726)	(1.386)	(1.513)
Job Satisfaction		-6.402*		0.552
		(3.330)		(5.987)
Workforce Age		0.217***		0.134
		(0.079)		(0.162)
Constant	6.488***	3.350	8.690***	2.892
	(0.737)	(4.080)	(1.420)	(8.341)
Observations	248	248	92	92
R^2	0.26	0.30	0.35	0.35

*** p<0.01, ** p<0.05, * p<0.1, robust standard errors in parentheses
OLS regression: Further controls: 2 firm size dummies and
11 industry dummies
Reference category: 0-99 employee firm in the food industry

Table 4.8: Performance-related pay and absenteeism

4.4 Conclusion

The aim of this study was to identify the relationship between incentive schemes and the level of cooperation among workers. We could make use of a large representative employer-employee survey, spanning a representative sample of firms from all industries which contains much more detailed information on the structure of incentive schemes as compared to data sets that have previously been used. Investigating this data set, we detected a positive relationship between the intensity of average team incentives in a firm and perceived helping efforts. We did not find similar effects for variable compensation based on company performance. This observation is well in line with what we expect from a standard agency model: Apparently, performance pay based on overall firm success is not sufficient to induce higher helping efforts as there is a large free rider problem which is much weaker when the performance of specific teams is measured. Moreover, our results indicate that higher individual performance pay has no negative consequences for helping efforts and that the positive effects of team incentives are not driven by self-selection. In line with these findings, we also found less absenteeism in firms providing stronger team incentives but not in firms using higher levels of individual performance pay.

All in all, our results strongly support the idea that team incentive schemes are a key component in a firm's incentive strategy and substantially affect the level of cooperation in organizations.

4.5 Appendix

Variable	*Description*	*Min*	*Max*	*Mean*	*Sd*	*N*
Individual PRP	% of total wage depending on individual performance	0	56	2.70	6.78	289
Team PRP	% of total wage depending on team performance	0	40	0.79	3.01	290
Firm PRP	% of total wage depending on firm performance	0	20	0.92	2.65	289
Hierarchical Levels	Number of hierarchical levels in the firm	1	9	2.5	1.14	295
Works Council	Works council in the firm (Yes/No)	0	1	0.60	0.49	294
Industry	12 industry dummies	1	12	-	-	305
Firm Size	3 firm size dummies: 0-99, 100-499 and $\geq$ 500 employees	1	3	-	-	305
Part-time Workers	Percentage of employees in part-time occupation	0	91.65	20.05	21.1	305
Wage Level	-1= below tariff, 0= equal to tariff, 1= above tariff	-1	1	-0.29	0.53	268
Gender Career	Female employees career support in the firm (Yes/No)	0	1	0.06	0.25	294
Trainings	Average number of trainings per employee	0	321	14.09	44.25	171
Workforce Change	-1= workforce reduction, 0 = no change, 1= workforce increase	-1	1	0.05	0.78	305
Workforce Age	Mean age of employees in the firm	28	50.63	40.17	4.00	305
Importance of Helping	% of workforce for which helping others with one's job is important	33	100	76.52	13.37	305
Job Satisfaction	% of workforce who are satisfied with their job	20	100	75.20	13.26	305

Table 4.9: Descriptive statistics of explanatory variables

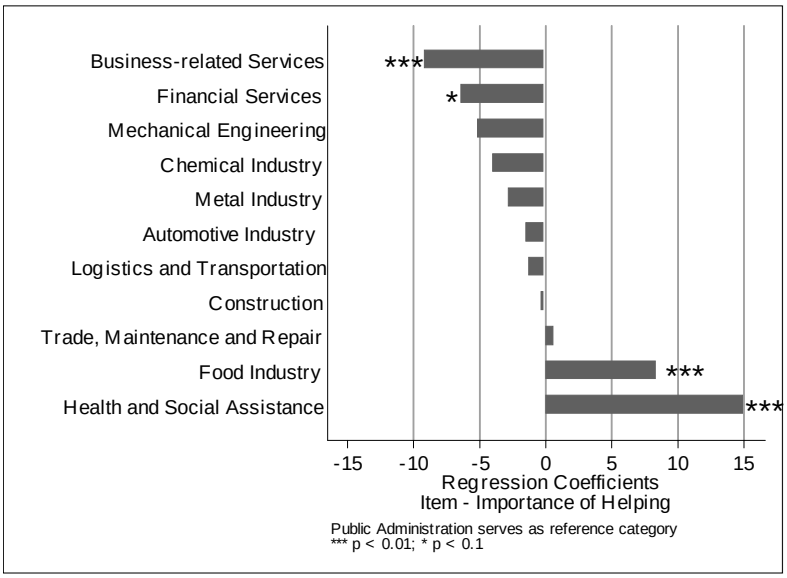

Figure 4.3: The fraction of cooperative employees across German industries

	All Firms				
Dependent Variable:	"You can count on the people to cooperate"				
	(1)	(2)	(3)	(4)	(5)
Individual PRP	0.0545	0.0153	0.0164	-0.0107	-0.261**
	(0.138)	(0.156)	(0.145)	(0.121)	(0.116)
Team PRP	0.626***	0.672***	0.652***	0.557***	0.665***
	(0.231)	(0.222)	(0.176)	(0.195)	(0.181)
Firm PRP	0.112	0.0697	-0.134	-0.131	-0.143
	(0.353)	(0.362)	(0.330)	(0.308)	(0.445)
Hierarchical Levels	3.346***	3.719***	3.787***	3.051***	2.505**
	(0.793)	(0.788)	(0.751)	(0.667)	(1.076)
Works Council	-3.931*	-4.606**	-3.186	-2.047	0.0705
	(2.311)	(2.305)	(2.370)	(2.197)	(2.987)
Part-time Workers	0.169**	0.127*	0.132*	0.0917	0.115
	(0.0773)	(0.0747)	(0.0718)	(0.0618)	(0.0894)
Wages below Tariff		-4.176	-4.975	-2.598	-8.164
		(3.985)	(3.842)	(3.012)	(7.400)
Wages above Tariff		-0.0531	-0.0325	0.229	2.178
		(2.144)	(2.087)	(1.839)	(2.855)
Workforce Reduction			-1.871	-0.665	-3.258
			(2.325)	(2.193)	(3.607)
Workforce Increase			6.832***	5.481**	7.546**
			(2.443)	(2.271)	(3.326)
Job Satisfaction				0.478***	0.408**
				(0.110)	(0.179)
Gender Career					-5.246
					(8.842)
Trainings					-0.011
					(0.028)
Constant	51.02***	51.03***	47.07***	12.92	11.37
	(5.274)	(5.257)	(5.656)	(9.220)	(15.39)
Observations	281	257	257	257	145
R^2	0.24	0.26	0.29	0.40	0.40

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

OLS regression: further controls: 11 industry and 2 firm size dummies

Reference category: 0-99 employee firm in the food industry

Table 4.10: Performance-related pay, HR policies and cooperation among workers

Chapter 5

Gender Differences in Risk Preferences among Workers and Managers - Field Evidence from Germany[1]

5.1 Introduction

Risk attitudes crucially affect behavior in various domains of life. The degree of risk aversion determines, for instance, entrepreneurship (Caliendo et al. (2011)), industry choice, portfolio choice (Dohmen et al. (2011)) and self-selection into payment schemes (Dohmen and Falk (2011)). Understanding the antecedents of individual risk aversion is therefore often necessary for understanding and predicting individual decision making.

Previous studies have frequently looked at gender as a central determinant of risk aversion. Based on behavioral risk measures involving real-stake lotteries or investment choices, recent experimental studies provide rather consistent evidence that women are less willing to take risks than men (see for instance Holt and Laury (2002), Eckel and Grossman (2002), Eckel and Grossman (2008a) or Croson and Gneezy (2009), Eckel and Grossman

[1] This chapter is based upon Berger (2011).

(2008b) and Charness and Gneezy (2010) for recent surveys). However, most of these experiments study gender effects among students or the "general population". Studies involving professionals and managers are less clear cut. Indeed, several studies find no systematic difference in investment behavior among professional male and female fund managers (e.g. Atkinson et al. (2003)) or among students who have undertaken formal management training (e.g. Johnson and Powell (1994)). In chapter 2.4 of their survey article Croson and Gneezy (2009) summarize the evidence on gender-specific risk differences among managers and professionals as follows: *"The conclusion is that gender differences in risk preferences among the general population do not extend to managers. This could be the result of selection; people that are more risk taking tend to choose managerial positions. While fewer women select these positions, those that do choose them have similar risk preferences as men. This result could also be an adaptive behavior to the requirements of the job."* (p. 6-7)

In this paper I test this *"important exception to the rule"* with two representative, yet independent surveys, the 2009 wave of the German Socio-Economic Panel (GSOEP) and a unique representative employer-employee matched survey data set containing 305 firms and more than 36,000 individual employee responses of the year 2006 (GPTW), the data set which has already been introduced in chapter 4. Both surveys not only contain items that assess individual risk preferences but also distinguish between employees working in non-managerial and different managerial positions. The benefit of analyzing the GPTW data set in addition to the GSOEP is that it allows me to control for selection and unobserved firm-fixed effects. Furthermore, the data set provides richer firm information, which I use to study additional determinants of risk aversion such as a company's incentive system.

The main result of the paper is that women are, at all hierarchical positions, significantly more risk averse than men. Furthermore, the gender risk gap is neither systematically lower nor systematically higher among managers than among workers. In line with the literature, managers are on average less risk averse than employees in non-management positions and top managers are less risk averse than managers in the lower or middle management. The

results are fairly similar across both data sets.

The remainder of the paper is organized as follows. I first test my research question with the GSOEP which I briefly explain and then analyze in the next chapter. Chapter 5.3 takes similar steps and tests my main research question using the GPTW data set. An additional chapter focuses on the interplay of gender, risk preferences and performance-related pay in some more detail, before a conclusion is presented in chapter 5.5.

5.2 GSOEP: Data, Methods and Results

The GSOEP is a longitudinal and representative survey containing detailed information of over 20,000 individuals in roughly 12,000 households in Germany. Individuals give information on socio-demographics, job and work related attributes and make subjective assessments of individual preferences, including risk, personality, satisfaction and the like.[2]

Key to my analysis are the subjective assessments of the individuals' general willingness to take risk. In particular, individuals are asked: *"How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?"* The question is to be answered on a 11-point Likert scale ranging from 0 *"unwilling to take risks"* to 10 *"fully prepared to take risks"*.[3] This validated survey measure is a reliable predictor of actual risk taking behavior (Dohmen et al. (2011), Hardeweg et al. (2011) and Ding et al. (2010)). The measure also exhibits test-re-test stability (Lönnqvist et al. (2011)) and even seems to dominate the popular real-stake risk elicitation method introduced by Holt and Laury (2002) (Hardeweg et al. (2011) and Lönnqvist et al. (2011)). In addition, two more context-specific but similarily worded risk measures concerning the willingness to take risks in one's occupation and in financial matters are investigated.[4]

[2] A detailed description of the data set can be found in Burkhauser and Wagner (1993) and Schupp and Wagner (2002)

[3] Translated from German.

[4] Dohmen et al. (2011) report that the general risk question turns out to be the best risk measure across different domains of risk taking while the more context-specific risk measures, also included in the GSOEP, have a higher predictive power for the particular

The 2009 wave contains a new variable, characterizing the individual's hierarchical position at work.[5] To be precise, every person was asked if he or she is in a leadership position, and if so, whether he or she is in a highly qualified specialist position (e.g. project head), in the lower management (e.g. group supervisor, section head), in the middle management (e.g. department head, regional director) or in the top management (e.g. executive board, business director, division manager). This information allows me to analyze gender differences in risk attitudes at the non-managerial level and at four different managerial levels.[6]

Figure 5.1 displays the distribution of the general willingness to take risks in the sample.[7] Survey answers are nicely distributed with the modal response being the mid point of the scale and roughly half of the distributional mass below it.

To provide a graphical illustration of my main research question, I classify an individual as "risk averse" if he or she ticked 5 or below on the 11-point scale. Arguably, the classification is arbitrary as I do not know if these individuals are in fact risk averse in the strict sense. However, this classification yields a proportion of risk averse individuals similar to previous estimations derived from real-stake lotteries.[8]

Figure 5.2 displays, separately for women and men, the proportion of "risk

risk context.

[5] This variable was previously only included in the 2007 wave of the GSOEP. However, in that year the risk questions were not included. I therefore restrict my analysis to the year of 2009 in which both key variables are available. For robustness checks, I took the risk attitudes from 2006 for the 2007 wave and re-ran all my analyses with pooled cross sections and random effects regressions for the years 2007 and 2009. The main results remain qualitatively very similar. Results are available on request.

[6] Fietze et al. (2010) also use this variable to study personality differences between female and male leaders in the 2007 wave of the GSOEP. They only briefly look at gender differences in risk attitudes and seem to find similar results as I do. However, they only use binary information to distinguish leaders from non-leaders and only look at sample means.

[7] I restrict the GSOEP sample to full or part-time employees. However, the results are not sensitive to this restriction.

[8] According to my classification roughly 76% of the sample is risk averse. In the seminal work of Holt and Laury (2002) 81% of subjects are risk averse. Using real-stake lotteries, Dohmen et al. (2011) estimate that 78% of the GSEOP sample is risk averse. In a supplementary analysis, Dohmen et al. (2011) use the same cut-off point to run binary regressions to explain risk aversion.

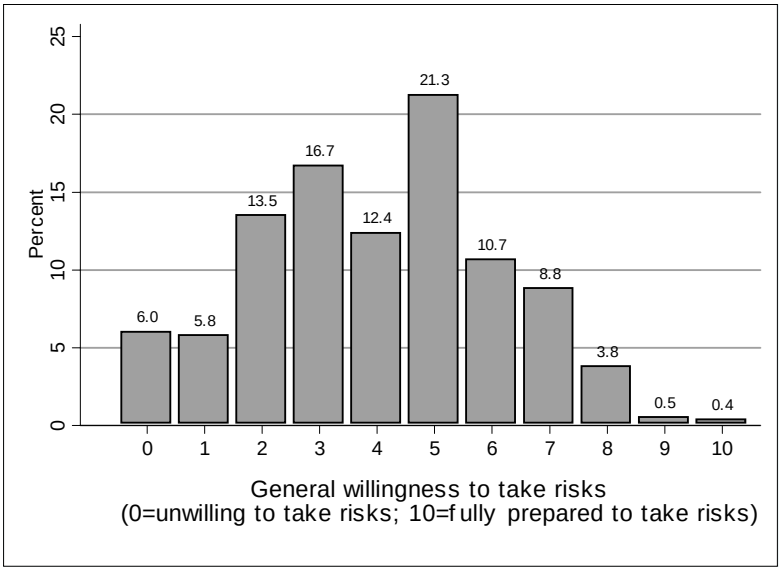

Figure 5.1: Distribution of the general willingness to take risks in the GSOEP 2009

averse" employees at all hierarchical levels and the 95% confidence intervals of the respective estimates.

First, the figure clearly shows that, on average, the share of risk averse women is substantially higher than the share of risk averse men on any given level. On each position, the share of women ticking one of the lower scale points is at least 10% points higher and this difference does not systematically vary in the hierarchical level.[9] Second, it seems that the general level of risk aversion decreases in the hierarchical level, i.e. managers are less risk averse than non-managerial workers. The share of risk averse women decreases from 84% in the non-managerial domain to 65% in the top management. For men, the respective numbers are 73% and 55%. Given the decrease in the absolute level of risk aversion among managers, the gender-specific risk difference even increases in relative terms when moving up the hierarchy. The displayed confidence intervals also suggest that the differences are significant.

Of course, the graphical representation of the main result may be incon-

[9]It seems that the difference get slightly larger for middle managers and slightly smaller for top managers.

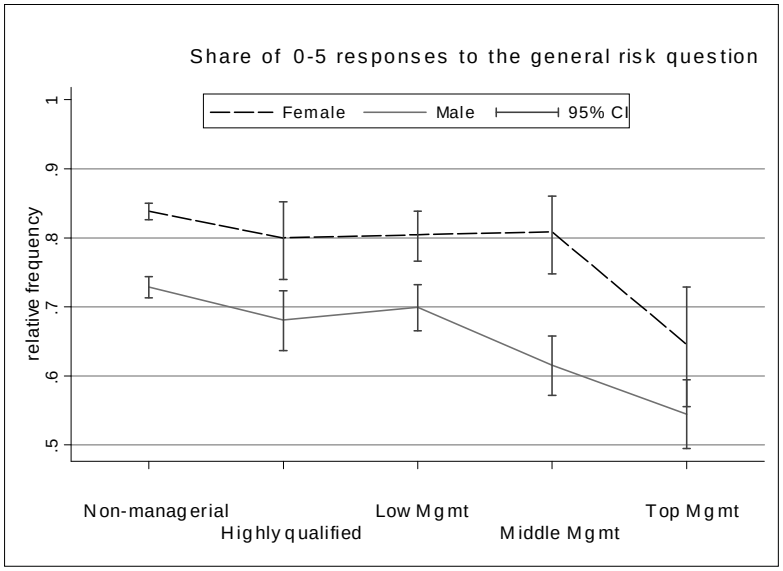

Figure 5.2: Gender-specific risk differences across hierarchical levels in the GSOEP - descriptives

clusive. The general level of risk aversion could, for instance, systematically differ across industries. As some industries are known to be male-dominated while others are more female-dominated, the observed differences may be driven by industry rather than gender effects. Other individual characteristics such as age, employment status or income may also be confounding factors which should be taken into account.[10] Similar to previous studies (e.g. Bell (2005)), female managers in the GSOEP are significantly younger and earn less money than their male counterparts. Given that risk aversion is expected to increase with age (e.g. Dohmen et al. (2011)), the risk gap among managers may turn out to be even larger than suggested by figure 5.2, once age is controlled for. In contrast, a higher income may allow individuals to become more risk seeking.[11] The persisting risk differences may

[10]Note that gender is, however, a perfectly exogeneous variable and thus less likely to be confounded by unobserved characteristics.

[11]Note that causality may also run the opposite way. Risk seeking behavior not only bears the risk of loosing more but also the chance of winning more. Since we do not observe individuals that were very unsuccessful, e.g. became unemployed or died, chances are that risky behavior determines higher incomes in the sample.

thus simply be attributed to the gender wage gap.

In order to isolate the average gender-specific risk difference across levels more precisely, I use a multiple regression approach that allows me to additionally control for age, education, income, marital status, origin, employment status, job category, firm size and industry. Descriptive statistics of these variables are provided in table 5.7. Due to the ordinal structure of the dependent variable, I run ordered probit regressions. I first regress the willingness to take risks on a female dummy and a set of control variables for all five job positions separately. The results are presented in table 5.1.

Dependent Variable:	"General willingness to take risks" (0=unwilling to take risks; 10=fully prepared to take risks)				
	Non-managerial	Highly qualified	Lower Mgmt	Middle Mgmt	Top Mgmt
	(1)	(2)	(3)	(4)	(5)
1 if Female	-0.285***	-0.383***	-0.295***	-0.436***	-0.217*
	(0.0331)	(0.108)	(0.0781)	(0.103)	(0.124)
Age (years)	-0.0133***	-0.0148***	-0.0156***	-0.00877*	-0.00627
	(0.00148)	(0.00522)	(0.00355)	(0.00515)	(0.00528)
Log Income	0.0620**	-0.0993	-0.0319	-0.0538	0.124*
	(0.0288)	(0.0980)	(0.0767)	(0.106)	(0.0736)
Observations	6,028	635	1,169	683	495
Pseudo R^2	0.02	0.02	0.02	0.03	0.03

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

Ordered probit regression: Firm controls: industry dummies (11), firm size dummies (3)
Individual controls: age, years of education, marital status dummies (3),
1 if German, 1 if full-time employee, 1 if east Germany, job category dummies (4)

Table 5.1: Gender-specific risk differences across hierarchical levels in the GSOEP - regressions

The results obtained from the regression analysis parallel the visual impression derived in figure 5.2. Even after controlling for various individual and firm characteristics women are, irrespective of the hierarchical level, significantly more risk averse than men. The female coefficients on the first three management levels are highly significant and at least the size of female coefficient of non-managerial employees. Computing the marginal effects of the estimates reveals that, for instance, women in the middle management are 3.3% more likely than men to consider themselves as "unwilling to take risks" (i.e. they tick 0 on the scale from 0-10). This difference is substantial given that only 3.7% of all middle managers give this response. In comparison, non-managerial women are 3.5% more likely to tick 0 than their male counterparts, while the sample average for this value is roughly 7% among all non-managerials. The female coefficient regarding the top management positions is only marginally significant and somewhat smaller than in the first column. Similar to previous findings, risk aversion decreases with age and is higher for employees with higher incomes.[12]

To test if the estimated female dummies differ significantly across levels, I interact gender and job level. The regression also includes four hierarchy dummies (with non-managerial workers being the reference category) to test if risk aversion decreases on higher levels. I run the regression on the general risk item and two context-specific risk measures, the willingness to take risk in once occupation and the willingness to take risk in financial matters. Table 5.2 shows the results for the variables of interest.

The highly significant female dummy resembles the gender-specific risk difference among non-managerial employees also obtained in column 1 of table 5.1. As predicted, the willingness to take risk increases in the hierarchical position (among males), indicated by the economically and statistically increasing coefficients from *Highly Qualified* to *Top Mgmt.*[13] Compared to male non-managerials, male employees in the lower management are 0.7%, male employees in the middle management are 2.6% and male top managers are

[12] As Dohmen et al. (2011) argue, age and income and other control variables included in the regression may be endogeneous to individual risk preferences. Including or dropping them from the analysis does not affect the gender estimate by much.

[13] The differences between the different levels are also significant.

Dependent Variable:	"Willingness to take risks" (0=unwilling to take risks; 10=fully prepared to take risks)		
	General Risk	Job Risk	Financial Risk
	(1)	(2)	(3)
1 if Female	-0.289***	-0.234***	-0.314***
	(0.0316)	(0.0320)	(0.0330)
1 if Highly Qualified	0.112**	0.127**	0.0684
	(0.0527)	(0.0533)	(0.0539)
1 if Lower Mgmt	0.0713*	0.178***	0.0928**
	(0.0432)	(0.0436)	(0.0442)
1 if Middle Mgmt	0.305***	0.418***	0.147***
	(0.0519)	(0.0523)	(0.0530)
1 if Top Mgmt	0.469***	0.527***	0.251***
	(0.0639)	(0.0643)	(0.0649)
Female × Highly Qualified	-0.0246	0.0319	-0.0847
	(0.0913)	(0.0926)	(0.0952)
Female × Lower Mgmt	0.0507	0.0427	-0.0812
	(0.0670)	(0.0676)	(0.0698)
Female × Middle Mgmt	-0.155*	-0.127	-0.0646
	(0.0910)	(0.0918)	(0.0938)
Female × Top Mgmt	-0.0426	-0.0420	-0.0738
	(0.110)	(0.111)	(0.113)
Observations	9,010	8,964	8,987
Pseudo R^2	0.02	0.03	0.03

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

Ordered probit regression: Firm controls: industry dummies (11), firm size dummies (3)

Individual controls: age, years of education, log income, marital status dummies (3),

1 if German, 1 if full-time employee, 1 if east Germany, job category dummies (4)

Table 5.2: Gender-specific risk differences across hierarchical levels in the GSOEP - interacted regressions

3.5% less likely to consider themselves as unwilling to takes risks. Second, the interaction between the female dummy and the hierarchical levels do neither yield sizeable nor significant differences confirming the absence of a decrease in gender-specific risk differences on managerial positions. Third, the results for the general risk question extend to the other two context-specific risk questions regarding financial matters and occupation, two risk domains which are of particular importance for managerial decision making.[14]

5.3 GPTW: Data, Methods and Results

Even though the GSOEP is a representative sample of the German population and includes both a validated measure of risk attitudes as well as detailed information on the employees managerial position, the observed gender difference among managers may result from a lack of proper controls. In particular, the risk difference may be driven by unobserved firm-specific characteristics forcing managers to be more or less risk averse at given managerial levels. If female managers worked in companies that generally induce lower levels of risk taking, treatment (female) and control group (males) would systematically differ with respect to company-specific effects. Since I do not observe males and females within the same firm, such arguments cannot be fully ruled out with the GSOEP analysis.

To address the problem of self-selection and unobserved firm-characteristics, I additionally explore a 2006 linked firm-worker survey conducted by the Great-Place-to-Work Institute and the German Federal Ministry of Labour and Social Affairs. The data set is representative for Germany and contains detailed information on 305 German firms employing a minimum of 20 workers. For each firm a management representative provided company-level information on organizational facts and HR instruments such as the structure of incentive systems. Most of this information is provided separately for managers and workers in each firm.[15] The management survey, for instance,

[14] In a recent AER paper Barseghyan et al. (2011) found that, in contrast to standard theory, risk preferences are not stable across different decision contexts.

[15] More specifically, answers were provided for employees in supervisory function and

includes detailed information on the structure of incentive pay. For both managerial and non-managerial employees the share of the average wage (in %) determined by performance-related pay (henceforth PRP) is known.[16]

In addition to the management survey, the data set includes a representative employee-survey yielding over 36,000 observations spread across the 305 firms. Employees provide bio-demographic information on age, tenure, gender, education and state if they work as a non-managerial, lower/middle managerial or top managerial employee. The data set thus allows me to observe males and females at the same hierarchical position within the same company. This within-firm variation should decrease problems that may arise when male and female employees self-select into systematically different work environments.

Most of the items contained in the employee survey aim to measure the general level of trust, pride, cooperation and leadership quality for the company as a whole. However, the survey also includes items on individual preferences. Employees rate their subjective importance of a high income, good development opportunities and job security. Even though the job security item may be more fuzzy than the validated risk measure in the GSOEP, it may still highly correlate with risk attitudes as preferences for security can be seen as the opposite of preferences for risk. The item is answered on a 5-point scale ranging from 1 "not important at all" to 5 "very important". A histogram of all employee answers is depicted in figure 5.3.

The distribution of the risk proxy is heavily skewed to the left, with over 80% of observations stating that job security is "very important" and almost everybody else stating that job security is "important" to them. To parallel the previous analytic steps, I create a binary variable taking on the value 1 if job security is very important and 0 otherwise. I then graph the distribution of this dummy, separately for women and men, on each hierarchical level while again including the estimates' respective 95% confidence intervals. The result is presented in figure 5.4.

Again we observe that the proportion of "risk averse" women is larger on

the largest group of non-managerial employees, i.e. the core occupational group.

[16] See chapter 4 for a more detailed analysis of performance-related pay components.

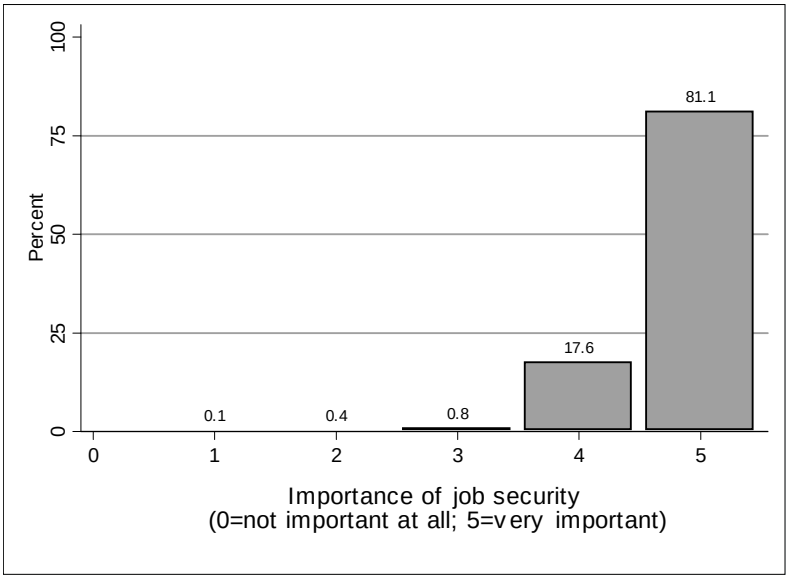

Figure 5.3: Distribution of importance of job security in the GPTW 2006

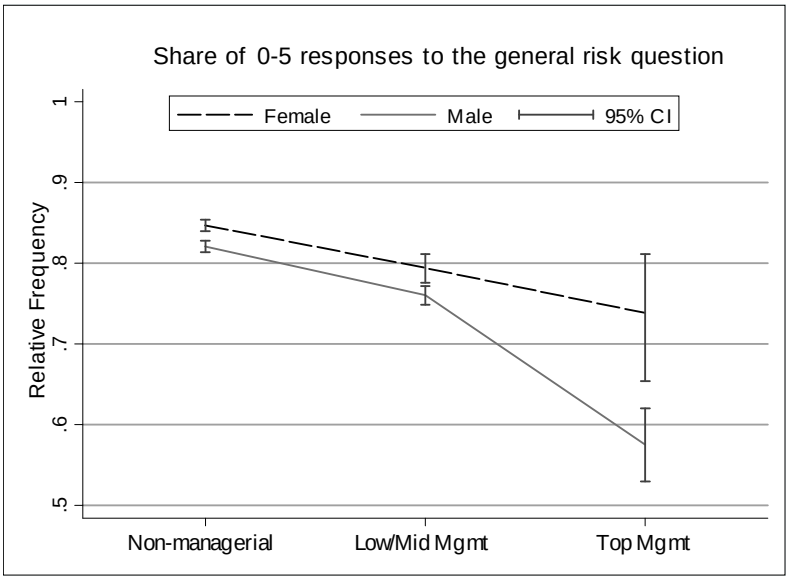

Figure 5.4: Gender-specific risk differences across hierarchical levels in the GTPW - descriptives

each hierarchical level. While the size of the differences seem to be smaller in absolute terms compared to the previous results, the 95% confidence intervals suggest that they are also significant. Again the general level of risk aversion decreases when climbing up the hierarchy, and again the gender-specific risk differences are, if at all, higher and not lower among managers.

Analogous to the previous chapter, I regress the risk aversion item on a female dummy and a set of standard individual and firm controls, separately for each given level. I also include a dummy which takes on the value 1 if wages on that particular level contain a performance-related pay component and zero otherwise. While agency theory predicts that risk averse agents will avoid wage uncertainty caused by incentive pay and prefer fixed wages (e.g. Prendergast (1999) and the references therein), empirical evidence on this topic is rather scarce.[17] To account for the fact that individuals are employed at the same firm, standard errors are clustered on the firm level. The left panel of table 5.3 applies an ordered probit regression on the job security item. The right panel reports marginal effects from simple probit models in which the latent variable is the propensity of perceiving job security as very important.

[17]Recent field evidence comes from Bellemare and Shearer (2010), Grund and Sliwka (2010) who find that risk averse workers are less likely to work under incentive pay. Dohmen and Falk (2011) provide clean experimental evidence that risk averse individuals are significantly less likely to self-select into piece rates or tournaments schemes.

Dependent Variable:	"Importance of job security" (0=not important at all; 5=very important)			"Importance of job security" 1 if very important		
	Ordered Probit			Probit (marginal effects)		
	Non-managerial	Lower & Middle Mgmt	Top Mgmt	Non-managerial	Lower & Middle Mgmt	Top Mgmt
	(1)	(2)	(3)	(4)	(5)	(6)
1 if Female	0.197***	0.124**	0.360**	0.0461***	0.0417***	0.107*
	(0.0306)	(0.0515)	(0.162)	(0.00776)	(0.0147)	(0.0588)
Age (years)	1.04e-05	0.00212	0.00930	-0.000157	0.000488	0.00335
	(0.00122)	(0.00204)	(0.00710)	(0.000297)	(0.000634)	(0.00287)
1 if PRP	-0.105*	-0.263***	-0.324**	-0.0246	-0.0759***	-0.131**
	(0.0606)	(0.0758)	(0.157)	(0.0155)	(0.0230)	(0.0617)
Observations	18,282	6,581	548	18,282	6,581	548
Pseudo R^2/ R^2	0.05	0.06	0.07	0.05	0.07	0.09

*** p < 0.01, ** p < 0.05, * p < 0.1, standard errors clustered on firms in parentheses

Firm controls: industry dummies (11), firm size dummies (3), 1 if works council

Individual controls: education dummies (6), 1 if full-time employee, 1 if non-physical work, 1 if employed in headquarter

Table 5.3: Gender-specific risk differences across hierarchical levels in the GPTW - regressions

Starting with the left panel, the results again match the graphical representation of the raw data. On each job level, women are significantly more risk averse, with the female coefficient being somewhat smaller for the lower and middle management but substantially larger for the top management level. Relating to the right panel, non-managerial female workers are roughly 5% more likely to consider job security as "very important" than their male counterparts. This difference remains almost identical for lower/middle managers but more than doubles to 11% among top managers.

Second, employees who receive PRP are less risk averse (less likely to consider job security as "very important"). This result seems to be particularly true for managerial employees.[18]

In the last step of the analysis I interact the job level variable with the female dummy to directly test for a decrease in the gender-specific risk difference on higher hierarchical levels. To rule out possible selection effects and the chance that unobserved firm or hierarchy inherent characteristics confound the relation between gender and risk preferences, I include firm-fixed effects in my regressions.[19]

[18] There are several possible explanations for this finding: First, in absolute terms (lower base wage) and in relative terms (lower % due to tariff systems) the PRP component for managerials is much higher and therefore more salient than for non-managerials. Second, bonus payments usually fluctuate more among managerials than among non-managerials. Third, managerials' PRP could in general by less affected by own work effort as bonus payments are usually based on firm-level figures.

[19] I did not include fixed-effects in the separate regressions because I wanted to investigate the link between the company's incentive system and risk aversion.

Dependent Variable:	"Importance of job security" (0=not important at all; 5=very important)		"Importance of job security" 1 if very important	
	Ordered Probit FE		OLS FE	
	(1)	(2)	(3)	(4)
1 if Female	0.177***	0.198***	0.0420***	0.0452***
	(0.0319)	(0.0336)	(0.00813)	(0.00853)
1 if Low or Middle Mgmt	-0.0526*	-0.0336	-0.0108	-0.00812
	(0.0281)	(0.0313)	(0.00699)	(0.00819)
1 if Top Mgmt	-0.365***	-0.357***	-0.111***	-0.114***
	(0.0597)	(0.0699)	(0.0207)	(0.0245)
Female × Low or Middle Mgmt		-0.0728		-0.0106
		(0.0518)		(0.0131)
Female × Top Mgmt		0.0916		0.0501
		(0.142)		(0.0469)
Firm Fixed Effects	Yes	Yes	Yes	Yes
Observations	26,380	25,411	26,380	25,411
Pseudo R^2	0.13	0.13	0.17	0.18

*** p < 0.01, ** p < 0.05, * p < 0.1, standard errors clustered on firms in parentheses
Individual controls: age, education dummies (3), 1 if full-time employee, 1 if non-physical work, 1 if employed in headquarter. Reference category: young male non-managerial employee with a degree from a professional school working a physical job

Table 5.4: Gender-specific risk differences across hierarchical levels in the GPTW - interacted regressions

The left panel of table 5.4 starts with the ordered probit regressions. Linear probability models explaining the risk dummy are added in the right panel to facilitate the economic interpretation of the effects.[20] Column 1 confirms an overall gender-specific risk difference and that risk aversion significantly decreases on higher hierarchical levels. The linear probability models yields that women are on average 4.5% more likely to assess job security as very important. Top managers are roughly 11% less likely than non-managerial workers to state that job security is very important to them. The second column adds the interaction between gender and job position and reveals that there are again no significant differences in the gender risk gap across levels. Also, the 5% increase in the gender gap among top managers is not significant. Moreover, the female coefficient, representing the gender risk gap among non-managerials, is virtually identical to the estimate in the previous table, suggesting that self-selection or firm unobservables do not drive the results.

While the firm-fixed effects are able to capture the general wage level in the firm, differences in individual wages cannot be controlled for in the GPTW data set. If we assume that female managers also earn less than male managers in the data set and that higher incomes are related with more risk seeking behavior, the estimated gender risk gap could be too large. Note, however, that not controlling for the individual wage level in the GSOEP analysis only slightly increases the gender coefficients. In addition, even though we do observe men and women in the same firm, it could still be the case that, on a given firm level, managerial jobs for women systematically differ from managerial jobs for men, e.g. they involve less risky decisions. Future studies should therefore try to analyze firm data with even more detailed information on hierarchical levels and job descriptions.

[20] Due to the interaction terms the estimation of marginal effects from probit models is not straighforward. However, the results for the non-interacted variables suggest that the linear probability model yields very similar results compared to the probit regression displayed in the previous table.

5.4 Application: Performance-related Pay, Risk and Female Managers

Currently, the underrepresentation of women in management positions and high-paid jobs is intensely debated in politics and economics. Table 5.5 displays the share of females across different managerial positions in Germany. While women make up roughly 50% of the workforce on the non-managerial level, the share of females decreases along the hierarchy. In the top management not even every 4^{th} position is held by a women. While the two different data sets provide rather similar results for the years 2006 and 2007, the GSOEP data suggests that the share of female managers slightly increased from 2007 to 2009.

	Share of Females (in %)		
Job Position	GSOEP 2009	GSOEP 2007	GPTW 2006
Non-managerials	52.6%	52.4%	49.5%
Highly Skilled	31.3%	27.3%	-
Lower Management	39.3%	36.0%	-
Middle Management	28.6%	27.9%	26.2%
Top Management	24.0%	20.1%	20.0%

Table 5.5: Share of women across hierarchical positions in Germany

Other sources reveal that in 2009 only 2.4% of all board members in the 500 largest firms were female.[21] While some countries (e.g. the Netherlands or Norway) have already taken actions to raise this number by introducing mandatory women quota, the source of female underrepresentation is still not fully understood by economists. Gender differences in risk preferences are considered one piece of the puzzle: Since managers' compensation usually depends to a much larger degree on bonuses and thus fluctuates more (see chapter 4), women may not want to apply for management positions because they tend to dislike wage uncertainty more than men.

In table 5.6 I test this conjecture with the GPTW data. Taking each firm as one observation, I use a simple OLS regression to explain the share

[21] Source: Hoppenstedt firm data base on June 2009.

of female managers in the firm (0-100%) by the existence and strength of managerial performance pay, holding constant other firm and industry characteristics. Following the line of thought described above, one would expect to see less female managers in firms that use performance-related pay for managers.

Dependent Variable	Share of female managers (in %)			
	(1)	(2)	(3)	(4)
1 if Manager PRP	-4.310**		-2.874	
	(1.937)		(1.867)	
1 if Manager PRP = 1-15%		-3.513*		-2.713
		(2.027)		(1.962)
1 if Manager PRP > 15%		-6.151***		-3.856*
		(2.364)		(2.301)
Control: Managers Risk Preferences	No	No	Yes	Yes
Observations	293	287	291	285
R^2	0.71	0.71	0.73	0.73

*** p < 0.01, ** p < 0.05, * p < 0.1, robust standard errors in parentheses
Firm controls: firmsize dummies (3), industry dummies (11), share of female non-managerials (in %), 1 if works council

Table 5.6: Share of female managers and performance-related pay

While the share of females among managers varies strongly across industries (from 7% in engineering to 66% in health and social work)[22], the share of female managers is indeed roughly 4 percentage points lower in companies in which managers receive performance-related pay. Given that in the collapsed data set women on average only make up 26% of all managers, the difference amounts to roughly 16% in relative terms. Column 2 suggests that the strength of incentives also matters. Compared to firms without managerial incentives, the share of female managers is even 6 percentage points lower when managerial PRP exceed 15% of the base wage.

A first indication that gender-specific differences in risk attitudes are at least one driver of this result is given by column (3) and (4). Here, I re-

[22]Table 5.5 displays the distribution of female managerial and non-managerial employees across all industries.

estimate the specifications in (1) and (2) but additionally control for the mean magnitude of managerial risk aversion. The coefficients for PRP decrease in economic terms and are not statistically significant at conventional levels. From this I infer that the correlation between the existence of performance-related pay and the share of females in the management is to some extent driven by the omitted variable "risk attitudes". Once risk preferences are controlled for, the existence of performance-related pay does no longer explain the percentage of female managers.

This parallels the experimental results by Dohmen and Falk (2011) who find that women are 23 percentage points less likely to self-select into variable pay schemes. However, the authors stress that this gender difference becomes much smaller and insignificant once controls for risk preferences are in place.

The example illustrates that current managerial compensation practices may also contribute to the relatively low share of female managers in Germany. Considering gender-based differentials in risk aversion when designing incentive schemes may be one option to make leadership positions more attractive for women.[23]

5.5 Conclusion

According to a recent literature overview by Croson and Gneezy (2009), managers are considered an exception to the general rule that women are more risk averse than men. The aim of the paper was to re-examine this conclusion. For this I analyze two independent and representative surveys of German employees. While the data of the German Socio-Economic Panel allow me to base my analysis on a validated survey measure of risk which has shown to be a reliable predictor of actual risk taking in real-stakes lotteries, the second data set includes observations of women and men within the same firm.

In both data sets the gender gap in risk attitudes remains roughly the same across levels, i.e. women are on all hierarchical levels more risk averse

[23]Currently, it seems that executive compensation is structured very similarly for men and women (see for instance Vieito and Khan (2010)).

than men. The pattern of risk aversion across levels and gender appears to be quite consistent across the two data sets. Moreover, as the analysis of the second data set shows, the results do not seem to be confounded by female managers systematically self-selecting into different firms than their male counterparts. While risk aversion generally decreases on higher hierarchical levels, the decrease is not larger for women and therefore does not offset the risk gap observed among non-managerials. This observation is in contrast to the conclusion drawn by Croson and Gneezy (2009) and the argument that after selection into management positions women and men do not differ with respect to risk attitudes.[24]

Apart from gender, the existence of performance-related pay is also significantly tied to the likelihood of observing a risk averse employee in the firm, affirming recent field evidence by Bellemare and Shearer (2010) and Grund and Sliwka (2010).

[24]The result is, however, in line with a working paper by Niessen and Ruenzi (2007) showing that female fund managers are more risk averse and pursue less extreme investment strategies than their male colleagues.

5.6 Appendix

Variable	*Description*	*Min*	*Max*	*Mean*	*Sd*	*N*
Willingness to take risks	Item: General willingness to take risks	0	10	4.03	2.12	9,010
Dummy "Risk Averse"	Dummy =1 if general willingness to take risks <5	0	1	0.76	0.43	9,010
Female	Dummy =1 if female	0	1	0.46	0.50	9,010
Full-time	Dummy =1 if fulltime	0	1	0.22	0.42	9,010
Non-managerial	Dummy =1 if non-managerial employee	0	1	0.67	0.47	9,010
Highly Qualified	Dummy =1 if highly qualified employee	0	1	0.07	0.26	9,010
Lower Mgmt	Dummy =1 if employee in lower management	0	1	0.13	0.34	9,010
Middle Mgmt	Dummy =1 if employee in middle management	0	1	0.08	0.26	9,010
Top Mgmt	Dummy =1 if employee in top management	0	1	0.05	0.23	9,010
Log Income	Logarithm of gross income	3.40	10.83	7.74	0.69	9,010
Age	Age of individual	18	84	44.18	10.68	9,010
Education	Years of education	7	18	12.91	2.76	9,010
German	Dummy =1 if German	0	1	0.95	0.22	9,010
East Germany	Dummy =1 if born in east Germany	0	1	0.23	0.42	9,010
Family status	4 dummies (single, married, divorced, widowed)	1	4	-	-	9,010
Job Categories	4 dummies (civil servants, employed, self-employed, trainee)	0	5	-	-	9,010
Firm Size	4 dummies (<20, 20-199, 200-1999, ≥2000)	1	4	-	-	9,010
Industry	12 industry dummies	1	12	-	-	9,010

Table 5.7: Descriptive statistics - GSOEP

Variable	*Description*	*Min*	*Max*	*Mean*	*Sd*	*N*
Importance of job security	Assessment of the importance of job security on a 5-point scale	1	5	4.79	.465	25,411
Dummy job security	Dummy =1 if job security is very important	0	1	0.41	.49	25,411
Female	Dummy =1 if employee is female	0	1	0.41	.49	25,411
Age	Age of the employee (midpoint of age categories)	19	$\geq$65	40.73	10.64	25,411
Fulltime	Dummy =1 if employee works fulltime	0	1	0.83	0.38	25,411
Non-managerial	Dummy =1 if employee works in a non-managerial job	0	1	0.72	0.45	25,411
Lower/middle mgmt	Dummy =1 if employee works in the lower or middle management	0	1	0.26	0.44	25,411
Top mgmt	Dummy =1 if employee works in the top management	0	1	0.02	0.15	25,411
Headquarter	Dummy =1 if employee works in the headquarter of the company	0	1	0.54	0.50	25,411
Non-physical work	Dummy =1 if employee works non-physical job	0	1	0.62	0.49	25,411
Education	4 educational dummies (no degree and three other degrees)	1	4	-	-	25,411
PRP	Dummy =1 if employee receives performance-related pay	0	1	0.46	0.49	25,411
Works council	Dummy =1 if company has a works council	0	1	0.80	0.40	25,411
Firm size	3 dummies (20-99, 100-499, $\geq$500)	1	3	-	-	25,411
Industry	12 industry dummies	1	12	-	-	25,411

Table 5.8: Descriptive statistics - GPTW

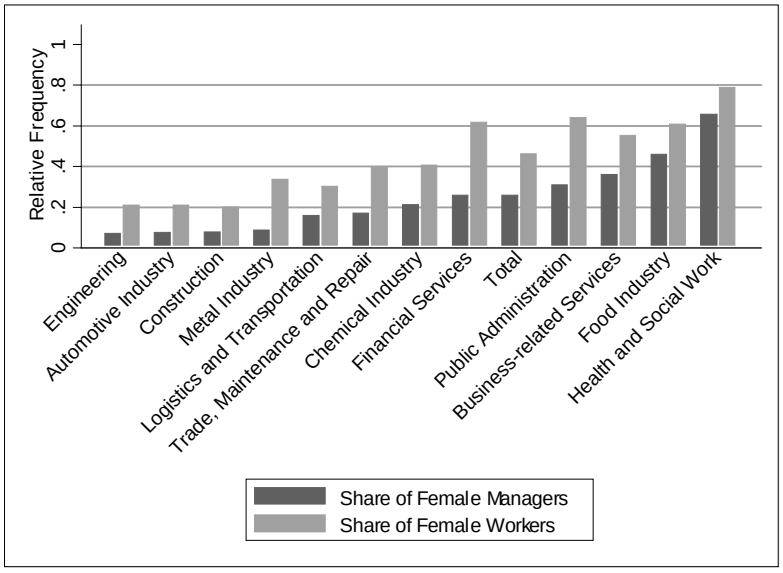

Figure 5.5: Share of female employees across German industries in the GPTW data set

Bibliography

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 8*, 1299–1324.

Ahn, T. S., I. Hwang, and M.-I. Kim (2010). The impact of performance measure discriminability on ratee incentives. *Accounting Review 85*, 389–417.

Akerlof, G. A. and J. L. Yellen (1988). Fairness and unemployment. *American Economic Review 78*, 44–49.

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

Alchian, A. A. and H. Demsetz (1972). Production, information costs, and economic organization. *American Economic Review 62*, 777–795.

Amegashie, J. A. (2009). American idol: Should it be a singing contest or a popularity contest? *Journal of Cultural Economics 33*, 265–277.

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

Atkinson, S. M., S. B. Baird, and M. B. Frye (2003). Do female mutual fund managers manage differently? *Journal of Financial Research 26*, 1–18.

Auriol, E., G. Friebel, and P. L. (2002). Career concerns in teams. *Journal of Labor Economics 20*, 289–307.

Bach, N., O. Guertler, and J. Prinz (2009). Incentive effects in tournaments with heterogeneous competitors - an analysis of the olympic rowing regatta in sydney 2000. *Management Revue. The International Review of Management Studies 20*, 239–253.

Baker, G., R. Gibbons, and K. J. Murphy (1994). Subjective performance measures in optimal incentive contracts. *Quarterly Journal of Economics 109*, 1125–56.

Bandiera, O., I. Barankay, and I. Rasul (2010a). Social incentives in the workplace. *Review of Economic Studies 77*, 417–458.

Bandiera, O., I. Barankay, and I. Rasul (2010b). Team incentives: Evidence from a firm level experiment. *Working Paper*.

Barseghyan, L., J. Prince, and J. C. Teitelbaum (2011). Are risk preferences stable across contexts? evidence from insurance data. *American Economic Review 101*, 591–631.

Bell, L. A. (2005). Women-led firms and the gender gap in top executive jobs. *IZA Discussion Paper 1689*.

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

Bellemare, C. and B. Shearer (2010). Sorting, incentives and risk preferences: Evidence from a field experiment. *Economics Letters 108*, 345–348.

Berger, J. (2011). Gender differences in risk preferences among workers and managers - field evidence from germany. *Working Paper*.

Berger, J., C. Harbring, and D. Sliwka (2010). Performance appraisals and the impact of forced distribution: An experimental investigation. *IZA Discussion Paper 5020*.

Berger, J., C. Herbertz, and D. Sliwka (2011). Incentives and cooperation in firms: Field evidence. *IZA Discussion Paper 5618*.

Berger, J. and P. Nieken (2010). Heterogeneous contestants and effort provision in tournaments - an empirical investigation with professional sports data. *SFB/TR 15 Discussion Paper No. 325*.

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

Bhattacherjee, D. (2005). The effects of group incentives in an indian firm: Evidence from payroll data. *Labour 19*, 147–173.

Blanco, M., D. Engelmann, and H.-T. Normann (2010). A within-subject analysis of other-regarding preferences. *Games & Economic Behavior 72*, 321–338.

Bol, J. C. (2011). The determinants and performance effects of managers' performance evaluation biases. *Accounting Review (forthcoming)*.

Bolton, G. E. and A. Ockenfels (2000). ERC - a theory of equity, reciprocity and competition. *American Economic Review 90*, 166–193.

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.

Brown, J. (2011). Qitters never win: The (adverse) incentive effects of competing with super stars. Working Paper.

Bull, C., A. Schotter, and K. Weigelt (1987). Tournaments and piece rates: An experimental study. *Journal of Political Economy 95*, 1–33.

Burkhauser, R. and G. Wagner (1993). The english language public use file of the german socio-economic panel. *Journal of Human Resources 28*, 429–433.

Burks, S., J. Carpenter, and L. Goette (2009). Performance pay and worker cooperation: Evidence from an artefactual field experiment. *Journal of Economic Behavior & Organization 70*, 458–469.

Cadsby, C. B., F. Song, and F. Tapon (2007). Sorting and incentive effects of pay-for-performance: An experimental investigation. *Academy of Management Journal 50*, 387–405.

Caliendo, M., F. Fossen, and A. Kritikos (2011). Personality characteristics and the decision to become and stay self-employed. *IZA Discussion Paper 5566*.

Caliendo, M. and D. Radic (2006). Ten do it better, do they? an empirical analysis of an old football myth. *IZA Discussion Paper No. 2158*.

Camerer, C. F. (1989). Does the basketball market believe in the 'hot hand,'? *The American Economic Review 79*, 1257–1261.

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

Charness, G. and U. Gneezy (2010). Strong evidence for gender differences in experimental investment. *Working Paper*.

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

Che, Y.-K. and S.-W. Yoo (2001). Optimal incentives for teams. *American Economic Review 91*, 525–54.

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

Cohn, A., E. Fehr, and L. Goette (2010). Fair wages and effort: Evidence from a field experiment. *Working Paper*.

Croson, R. and U. Gneezy (2009). Gender differences in preferences. *Journal of Economic Literature 47*, 448–474.

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

Deutscher, C., B. Frick, O. Gürtler, and J. Prinz (2009). Sabotage in heterogeneous tournaments: A field study. *Working Paper*.

Ding, X., J. Hartog, and Y. Sun (2010). Can we measure individual risk attitudes in a survey? *IZA Discussion Paper No. 4807*.

Dohmen, T. and A. Falk (2011). Performance pay and multi-dimensional sorting: Productivity, preferences and gender. *American Economic Review 101*, 556–590.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association 9*, 522–550.

Dohmen, T. J. (2008). The influence of social forces: Evidence from the behavior of football referees. *Economic Inquiry 46*, 411–424.

Drago, R. and G. T. Garvey (1998). Incentives for helping on the job: Theory and evidence. *Journal of Labor Economics 16*, 1–25.

Dufwenberg, M. and G. Kirchsteiger (2004). A theory of sequential reciprocity. *Games & Economic Behavior 47*, 268–298.

Dur, R. and J. Sol (2010). Social interaction, co-worker altruism, and incentives. *Games & Economic Behavior 69*, 293–301.

Eckel, C. C. and P. J. Grossman (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior 23*, 281 – 295.

Eckel, C. C. and P. J. Grossman (2008a). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization 68*, 1–17.

Eckel, C. C. and P. J. Grossman (2008b). Men, women and risk aversion: Experimental evidence. In C. R. Plott and V. L. Smith (Eds.), *Handbook of experimental economics results*, pp. 1061–1073. Elsevier.

Ehrenberg, R. G. and M. L. Bognanno (1990). Do tournaments have incentive effects? *Journal of Political Economy 98*, 1307–1324.

Encinosa, W., M. Gaynor, and J. B. Rebitzer (2007). The sociology of groups and the economics of incentives: Theory and evidence on compensation systems. *Journal of Economic Behavior & Organization 62*, 187–214.

Engellandt, A. and R. T. Riphahn (2011). Evidence on incentive effects of subjective performance evaluations. *Industrial & Labor Relations Review 64*, 241–257.

Eriksson, T. and M. C. Villeval (2008). Performance-pay, sorting and social motivation. *Journal of Economic Behavior & Organization 68*, 412–421.

Falk, A. and U. Fischbacher (2006). A theory of reciprocity. *Games & Economic Behavior 54*, 293–315.

Falk, A. and A. Ichino (2006). Clean evidence on peer effects. *Journal of Labor Economics 24*, 39–57.

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance 25*, 383–417.

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.

Fershtman, C. and U. Gneezy (2011). The trade-off between performance and quitting in high-power tournaments. *Journal of the European Economic Association 9*, 318–336.

Fietze, S., E. Holst, and V. Tobsch (2010). Germany's next top manager: Does personality explain the gender career gap? *IZA Discussion Paper 5110*.

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

Forrest, D. J., J. Goddard, and R. Simmons (2005). Odds-setters as forecasters: The case of english football. *International Journal of Forecasting 21*, 551–564.

Frick, B., O. Gürtler, and J. Prinz (2008). Anreize in turnieren mit heterogenen teilnehmern: Eine empirische untersuchung mit daten aus der fussball-bundesliga. *Zeitschrift für betriebswirtschaftliche Forschung 60*, 385–405.

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*, 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.

Grund, C. and D. Sliwka (2010). Evidence on performance pay and risk aversion. *Economics Letters 106*, 8–11.

Hamilton, B. H., J. A. Nickerson, and O. Owan (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy 111*, 465–497.

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*, 923–951.

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

Harbring, C., B. Irlenbusch, M. Kräkel, and R. Selten (2007). Sabotage in corporate contests - an experimental analysis. *International Journal of the Economics of Business 14*, 367–392.

Harbring, C. and G. Luenser (2008). On the competition of asymmetric agents. *German Economic Review 9*, 373–395.

Hardeweg, B., L. Menkhoff, and H. Waibel (2011). Experimentally - validated survey evidence on individual risk attitudes in rural thailand. *Discussion Paper*.

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 8*, 817–837.

Heywood, J. S., U. Jirjahn, and G. Tsertsvadze (2005). Getting along with colleagues - does profit sharing help or hurt? *Kyklos 58*, 557–573.

Holmström, B. (1982). Moral hazard in teams. *Bell Journal of Economics 13*, 324–340.

Holmström, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership and job design. *Journal of Law, Economics & Organization 7*, 24–52.

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American Economic Review 92*, 1644–1655.

Ichino, A. and G. Maggi (2000). Work environment and individual background: Explaining regional shirking differentials in a large italian firm. *Quarterly Journal of Economics 115*, 1057–1090.

Itoh, H. (1991). Incentives to help in multi-agent situations. *Econometrica 3*, 611–636.

Itoh, H. (1992). Cooperation in hierachical organizations: An incentive perspective. *Journal of Law, Economiocs, & Organization 8*, 321–345.

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*, 905–925.

Johnson, J. and P. Powell (1994). Decision making, risk and gender: Are managers different? *British Journal of Management 5*, 123–138.

Jones, D. C., P. Kalmu, and A. Kauhanen (2010). Teams, incentive pay, and productive efficiency: Evidence form a food-processing plant. *Industrial Labor Relations Review 63*, 606–626.

Jones, D. C. and T. Kato (1995). The productivity effects of employee stock-ownership plans and bonuses: Evidence from japanese panel data. *American Economic Review 85*, 391–414.

Kampkötter, P. and D. Sliwka (2011). Differentiation and performance - an empiricial investigation on the incentive effects of bonus plans. *mimeo*.

Kandel, E. and E. P. Lazear (1992). Peer pressure and partnerships. *Journal of Political Economy 100*, 801–17.

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

Knez, M. and D. Simester (2001). Firm-wide incentives and mutual monitoring at continental airlines. *Journal of Labor Economics 19*, 743–772.

Knoeber, C. and W. Thurman (1994). Testing the theory of tournaments: An empirical analysis of broiler production. *Journal of Labor Economics 12*, 155–179.

Konrad, K. (2009). *Strategy and Dynamics in Contests*. Oxford Business Press.

Kräkel, M. and D. Sliwka (2004). Risk taking in asymmetric tournaments. *German Economic Review 5*, 103–116.

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. (1989). Pay equality and industrial politics. *Journal of Political Economy 97*, 561–80.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review 90*, 1346–62.

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

Lazear, E. P. and K. L. Shaw (2007). Personnel economics: The economist's view of human resources. *The Journal of Economic Perspectives 21*, 91–114.

Levitt, S. D. (2004). Why are gambling markets organised so differently from financial markets?. *Economic Journal 114*, 223–246.

Lynch, J. (2005). The effort effects of prizes in the second half of tournaments. *Journal of Economic Behavior & Organization 57*, 115–129.

Lönnqvist, J.-E., M. Verkasalo, G. Walkowitz, and P. C. Wichard (2011). Measuring individual risk attitudes in the lab: Task or ask?: An empirical comparison. *SOEPpapers No. 370*.

Mas, A. and E. Moretti (2009). Peers at work. *American Economic Review 99*, 112–145.

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

Mohnen, A., K. Pokorny, and D. Sliwka (2008). Transparency, inequity aversion, and the dynamics of peer pressure in teams: Theory and evidence. *Journal of Labor Economics 26*, 693–720.

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.

Newhouse, J. (1973). The economics of group practice. *Journal of Human Resources 8*, 37–56.

Nieken, P. and M. Stegh (2010). Incentive effects in asymmetric tournaments - empirical evidence from the german hockey league. SFB TR 15 Discussion Paper No. 305.

Niessen, A. and S. Ruenzi (2007). Sex matters: Gender differences in a professional setting. *CFR-Working Paper No. 06-01*.

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*, 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.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review 83*, 1281–1302.

Rajan, M. V. and S. Reichelstein (2006). Subjective performance indicators and discretionary bonus pools. *Journal of Accounting Research 44*, 585–618.

Roman, F. J. (2009). An analysis of changes to a team-based incentive plan and its effects on productivity, product quality, and absenteeism. *Accounting, Organizations & Society 34*, 589–618.

Rosen, S. (1986). Prizes and incentives in elimination tournaments. *American Economic Review 76*, 701–715.

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

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *Quarterly Journal of Economics 116*, 681–704.

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.

Schupp, J. and G. G. Wagner (2002). Maintenance of and innovation in longterm panel studies: The case of the german socio-economic panel (gsoep). *DIW Discussion Paper No. 276*.

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.

Sunde, U. (2009). Heterogeneity and performance in tournaments: a test for incentive effects using professional tennis data. *Applied Economics 41*, 3199–3208.

Szymanski, S. (2003). The economic design of sporting contests. *Journal of Economic Literature 41*, 1137–1187.

Taylor, J. (2003). Risk-taking behavior in mutual fund tournaments. *Journal of Economic Behavior & Organization 50*, 373–383.

van Dijk, F., J. Sonnemans, and F. van Winden (2001). Incentive systems in a real effort experiment. *European Economic Review 45*, 187–214.

Vieito, J. P. and W. Khan (2010). Executive compensation and gender: S&p 1500 listed firms. *Journal of Economics & Finance*, 1–29.

Weigelt, K., J. Dukerich, and A. Schotter (1989). Reactions to discrimination in an incentive pay compensation scheme: A game-theoretic approach. *Organizational Behavior and Human Decision Processes 44*, 26–44.

Woodland, L. M. and B. M. Woodland (1994). Market efficiency and the favorite-longshot bias: The baseball betting market. *The Journal of Finance 49*, 269–279.

Wärneryd, K. (2000). In defense of lawyers: Moral hazard as an aid to cooperation. *Games & Economic Behavior 33*, 145–158.