The Evaluation of Human Resource Management Practices - Empirical Studies on Subjective Performance Evaluation, Target Agreements, and Intra-Firm Trainings

Inauguraldissertation

zur

Erlangung des Doktorgrades

der

Wirtschafts- und Sozialwissenschaftlichen Fakultät

der

Universität zu Köln

2011

vorgelegt von

Diplom-Kauffrau Kathrin Anne Breuer

aus

Mannheim

Referent:	Professor Dr. Dirk Sliwka
Korreferent:	Professor Dr. Ludwig Kuntz
Tag der Promotion:	16.12.2011

Für meine Familie

Danksagung

Die vorliegende Arbeit wäre nicht ohne die Unterstützung wichtiger Menschen entstanden, bei denen ich mich an dieser Stelle von ganzem Herzen bedanken möchte. Mein größter Dank gilt meinen Eltern und meinem Zwillingsbruder Timo, dafür dass sie immer hinter mir stehen und an mich glauben. Durch ihre stetige (telefonische und physische) Unterstützung geben sie mir großen familiären Rückhalt, der mir sehr viel bedeutet.

Ein ganz besonderer Dank geht an meinen Doktorvater Dirk Sliwka für seine außergewöhnliche Betreuung, die geprägt ist von fachlicher Kompetenz, ständiger Verfügbarkeit und Geduld. Ich bin sehr dankbar, dass ich in den letzten Jahren so viel von ihm lernen durfte. Zudem danke ich Prof. Ludwig Kuntz für die Übernahmen des Koreferats.

Des Weiteren möchte ich mich natürlich bei meinen lieben Kollegen bedanken. Allen voran danke ich meinem Ko-Autor Patrick Kampkötter (danke fürs Last-Minute Korrekturlesen), Claus Herbertz, Johannes Berger, meiner Ko-Autorin Petra Nieken, Torsten Biemann, Nannan Zhou, Anastasia Danilov, Christine Harbring, Bernd Irlenbusch und Felix Kölle für die gute Zusammenarbeit, die positive Arbeitsatmosphäre und die unterhaltsamen Flurgespräche. Sie sind in den letzten Jahren zu Freunden geworden, die ich nicht mehr missen möchte. Ein großer Dank geht natürlich auch an meinen Ko-Autor Jan-Hendrik Zimmermann, der trotz seines Berufseinstiegs immer für mich verfügbar war. Ich danke auch unserer Sekretärin Beate Ommer und allen studentischen Hilfskräften für die Hilfsbereitschaft und Unterstützung bei Recherchen, Befragungen und Experimenten: Ursula Schuh, Isabella Cetnarowksi, Katharina Laske, Philip Arminski, Alexander Creson, Christian Ruppert und Tobias Hinze. Weiterhin danke ich auch den ehemaligen

Hiwis und vor allem Andreas Staffeldt für sein Engagement bei der Durchführung von Experimenten.

Weiterhin möchte ich auch meinen Kollegen vom Graduiertenkolleg Risikomanagement danken, mit denen ich die ersten einerthalb Jahre meiner Promotionszeit verbracht habe: Martin Ruppert, Frowin Schulz, Tobias Wickern, Christof Wiechers. Ein besonderer Dank geht an meine liebe Büronachbarin Miriam Breunsbach für die schöne Zeit im kleinsten Büro des Kollegs. Auch meinen Kollegen der Universität Aarhus möchte ich ein herzliches Dankeschön aussprechen. Ich danke vor allem Nina Smith, Julia Nafziger und Alexander Koch für Ihre uneingeschränkte Gastfreundlichkeit und die hilfreichen Kommentare zu meiner Forschungsarbeit.

Ein sehr großes Dankeschön geht an meine lieben Freunde in Köln und Speyer für ihr offenes Ohr und die schönen gemeinsamen Stunden, die meine Freizeit stets "versüßt" haben. Besonders danken möchte ich meiner lieben Freundin Lena, dafür dass sie in guten wie in schlechten Zeiten immer für mich da war und den ein oder anderen Feierabend mit mir zelebriert hat. Ich danke auch der Familie Marggraf, Angelika, Dieter, Sabrina, Jonas, Laya und Lorenz für ihr stetes Interesse an meiner Arbeit und dem Daumendrücken bei Vorträgen.

Mein letzter und wichtigster Dank geht an meinen Freund Nikolas. Sein Glauben an mich und diese Arbeit haben mir von Tag zu Tag neue Kraft gegeben. Mit großer Geduld war er immer für mich da und hat mir vor allem in schwierigen Phasen mit aufbauenden Worten zur Seite gestanden. Ich bin sehr glücklich, dass es ihn gibt.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Motivation, effort and decisions of individuals are a major source of economic success in today's organizations. In order to effectively manage and lead individuals, organizations need to apply well-designed human resource practices. These enable firms to assign employees to the right jobs, to reduce employee turnover, to provide incentives to the workforce and to facilitate the recruitment of productive employees. As the human-resource related cost is the largest component of all costs in today's corporations, "one can hardly overestimate the importance of understanding better how organizations (...) manage their employees" (Lazear and Gibbs (2009)).

This thesis aims to contribute to the understanding of human resource instruments that are relevant in practice and discussed in the personnel economics and management literature. Four papers are presented which empirically investigate the use and the effects of three practices: the system of performance appraisal, the use of target agreements and intra-firm trainings. In the following, we will motivate the research on these practices, briefly describe the results and highlight how they are related to each other.

One major challenge for fulfilling the firm's organizational strategy is setting the right incentives for employees to induce optimal efforts. Incentives include remuneration systems such as individual incentive pay, promotions and career advancement. They are designed to overcome the conflict of interest between firms and employees which is that firms aim at maximizing

worker effort but effort is costly for employees. Incentive schemes could solve the conflict if employee effort would be perfectly observable and contractible. As this is typically not the case, firms use performance measures as a proxy for employee effort. But the use of performance measures, such as sales figures, adds noise to the process and increases uncertainty for employees. Since people are generally risk-averse, uncertainty reduces their motivation to exert higher effort. Hence, measurement of performance is a crucial part of incentive setting.

In many jobs there are not even objective performance measures available that capture employee performance, which is why firms often rely on subjective performance evaluations by supervisors (Prendergast (1999)). But these evaluations add supervisor discretion which might lead to biases in evaluations and dilute the incentive setting. When high performers are not adequately rewarded and low performers not adequately punished, the incentive to exert effort is reduced. There are two major evaluation biases discussed in the literature. First, supervisors tend to give inflated performance evaluations that are too positive relative to the true performance of employees (leniency bias). Secondly, there is a tendency to compress evaluations to the middle of the rating scale, so that the differentiation in grades is lower compared to the actual performance distribution of employees (centrality bias) (see Landy and Farr (1980), Murphy and Cleveland (1995) for an overview).

In the psychological and economic literature, several reasons for biased evaluations by supervisors are discussed. With lenient performance evaluations, supervisors may want to avoid conflicts with their subordinates (Napier and Latham (1986)) or agency costs for diverging beliefs about what constitutes good performance (MacLeod (2003)). Likewise, supervisors may want to signal their outstanding leadership competencies by assigning high grades to all of their employees. Moreover, positive reciprocity of employees for good evaluations may be expected as proposed in the efficiency wage literature (Akerlof (1982), see Fehr et al. (1993) for experimental evidence of gift exchange). Compressed ratings might occur due to supervisor's beliefs to have insufficient information about subordinates' performance. Instead of reward-

ing or punishing the wrong employees, they will tend to assign everyone the same grade. In addition, it is costly for supervisors to observe performance differences of subordinates and hence differentiate. However, one of the most prominent reasons for evaluation bias is that supervisors may have personal preferences for certain employees and favour them in evaluations (Prendergast and Topel (1996)). The degree of favoritism in evaluations might be determined by the social tie between supervisor and employee.

The second chapter of this thesis investigates whether social proximity between supervisor and employees may cause supervisors to give inflated ratings. Supervisors generally face a trade-off between evaluating accurately and caring for the well-being of their subordinates (Prendergast and Topel (1996)). The extent to which they care for the subordinate's well-being may depend on the social relation to the employee which is what we address in this study. Although the question is highly relevant with regard to incentive setting in organizations, empirical studies have been sparse, mainly because of limited access to meaningful data sets. With the study, we hence complement the empirical research on biases in performance appraisals.

Our study uses 4-years of personnel records of a call center subsidiary located in Germany. The unique feature of the data is that we are able to observe the performance evaluations by supervisors in addition to objective performance measures of the call center agents. Observing the "true" performance of agents allows us to identify potential biases in the evaluations. Social proximity is measured by the size of a work team and a variable indicating if an employee has worked for the supervisor before. For the analysis we apply fixed effects regressions and thus make use of frequent changes of agents between departments to eliminate unobserved heterogeneity of individuals. While controlling for the objective performance measures, we find that employees working in smaller teams are assessed more leniently than employees working in larger teams. Furthermore, employees who have worked for the same supervisor before receive better grades than employees who have not. In the analysis, learning effects of agents are controlled for by including firm tenure in the regressions. The results of the study are highly consequential when setting up performance evaluations in organizations as biases

may affect work morale and reduce performance (Bol (2011), Berger et al. (2010)). Firm strategies to mitigate the problem of biased evaluations are discussed in the conclusion of chapter 2.

In the third chapter, our focus turns to the determinants of centrality bias and differentiation in evaluations. Besides social interaction, also supervisor characteristics may influence how evaluations are made. Men and women may possibly differ in how they evaluate subordinates. Chapter 3 analyzes whether women evaluate in a more differentiated manner than men do. Assuming that higher rating differentiation is a leadership competence, we may contribute to the question of gender differences in leadership qualities. This question is of special interest in the recent debate of why women remain still rare in leadership positions. Based on former evidence on gender differences in social preferences and personality (Croson and Gneezy (2009)), we derive several arguments why women may differentiate in other ways than men.

Our empirical analysis is based on personnel records of a large, multinational company based in Germany for the years 2006-2008. We observe the results of subjective performance evaluations for all employees in Germany, yielding about $13,000$ employee-year observations. To every employee, we can match the direct supervisor who conducts the evaluations. We consider all employees that are working under the same supervisor as work unit. Two variables measure the degree of differentiation in grades: The standard deviation of grades and the percentage of poor grades per unit. With the aggregated data set, on work unit level, it is shown that women differentiate to a significantly larger extent between their subordinates than men do. A further analysis shows that female supervisors also give poor grades more frequently. Given an often reported reluctance of supervisors to identify low performers, women seem to overcome this task better than men. The results are confirmed by a further analysis that allows to eliminate time-constant unobserved heterogeneity between units. Therefore the data set is rearranged so that a work unit is defined based on a group of employees that worked together over the three years. The new data set allows to use gender of the supervisor as a time-varying work unit variable. Hence, we can use fixed effects regressions through which an effect is only identified when

supervisors change between units. In the data, we additionally observe information on the leadership evaluations that supervisors receive. Women score significantly higher in the leadership dimension "Leading People" which is consistent with the former result. Overall, the analyses suggests that, on average, female supervisors show good leadership abilities by evaluating in a more differentiated manner than men do. Several reasons for this observation are discussed. Besides gender differences in social preferences, a further reason might be a selection of very competent women in leadership positions. This suggests that different standards to become a leader might be applied for women compared to men which is further addressed in the conclusion of the chapter.

In practice, performance evaluations are often combined with target agreements serving as a performance standard against which employees' performance is evaluated (Murphy (2001)). Murphy (2001) emphasizes the relevance of these standards for bonus payments: "Bonuses are usually not, in practice, based strictly on a performance measure, but rather on performance measured relative to a performance standard" (Murphy (2001), p. 246). The setting of goals or standards for incentive schemes is a widely used practice in firms and recent studies have shown renewed interest in the process of goal setting (Anderson et al. (2010), Bol et al. (2010), Koch and Nafziger (2011)). Research on goals and targets is originally based on the goal-setting theory by Latham and Locke (1990) showing that goals have a motivational effect on people when they are set in a sufficiently specific and challenging manner. Economic arguments also suggest that using standards can lead to a higher pay-performance sensitivity (Murphy (2001)). While most of the empirical research has looked at the effects of target setting for individuals, empirical evidence on the use and the performance effect of target setting in firms is scarce.

The fourth chapter investigates which firms use target agreements for employees in Germany and whether the introduction of target agreements leads to an increase in overall firm performance. By deriving hypotheses about which firms will mostly benefit from target settings, the determinants of firms using this practice are investigated. We make use of a representa-

tive firm-level data set from the Federal Employment Agency (BA) at the Institute for Employment Research (IAB) in Germany. In 2005 and 2007, establishments were asked about the use of target agreements for employees. We find that qualification and tenure of the workforce are important determinants for the use of this practice as the cost of reduced effort is higher for qualified employees and longer-tenured employees require less guidance through specific goals. In addition, firms that have recently undergone reorganizations are more likely to use target settings than those who have not. Surprisingly, we do not find a firm size effect on the probability of using target agreements which has been expected due to economies of scale for the use of costly human resource practices. Unionization in firms shows a positive relation with using target agreements for employees while the reverse has been expected because of a possible influence of target agreements on payment schemes. For the analysis of a performance effect we make use of a first-differencing approach to eliminate unobserved heterogeneity between firms. The influence of time-constant factors such as the general growth potential of a firm in the observed years can thus be eliminated. We can further control for observed differences in firm size or workforce characteristics between the years. By comparing firms in a regression analysis that introduced target settings between 2005 and 2007 to those that have not, we find a substantial positive effect of the introduction on sales growth of 5.7%. In a robustness check, we additionally include a variable for having introduced subjective performance evaluation to rule out that the result is driven by the implementation of a formal evaluation system. The effects remain robust leading to the conclusion that firms seem to highly benefit from target setting for employees.

Besides the importance of performance evaluations for incentive schemes in organizations (Prendergast (1999)), evaluations serve as a feedback device through which training needs of employees may be identified (Murphy and Cleveland (1995)). Intra-firm training is the third human-resource related practice which is investigated in this thesis. Training investments generally aim to increase employees' human capital to make them more productive. Therefore, many studies have focused on identifying the productivity effects

of training by looking at individual wages (Bartel (2000), Dearden et al. (2006)) or firm-level productivity such as sales or value added (i.a. Black and Lynch (1996)). In the analysis of chapter 5, we address whether investment in training may lead to behavioral responses of employees. In particular, the effects of training on absenteeism and turnover are investigated based on a single-firm data set. Whereas human capital theory would expect a permanent decrease of absenteeism due to increased opportunity costs of leisure time, we raise the argument that employees might perceive firm-provided training as a gift (similar to the logic of efficiency wage theory by Akerlof (1982)) and positively reciprocate with lower absence. Economic theory further expects an increased turnover for employees that are trained in general skills (like in the firm we observe) because of their enhanced value for competitive firms. Here again, the argument is raised that training may be perceived as a gift or a signal of employer's confidence in the employee which may lead to the opposite effect and increase employee loyalty. In contrast to economic predictions, investments in general training may hence even help to reduce employee turnover.

The empirical analysis of chapter 5 is based on personnel records of a large, multinational company for the years $2006-2008$, similar to those used in chapter 3. The focus is partially on non-managerial employees because only for them we observe absence hours. Besides having detailed information on training participation, we can track whether employees voluntarily left the company in the year 2007 or 2008. One major issue discussed in the training literature is the endogeneity of being selected into training. Our empirical approach aims at addressing this in several ways. A panel approach is conducted for the analysis of absenteeism to eliminate time-invariant differences between employees such as general motivation or ability. Furthermore, we conduct propensity score matching by which an optimal control group is computed that is most similar to the group that received training. By comparing the outcome variable between the two groups, an average treatment effect on the treated (ATT) can be derived. Results of the regression analyses indicate that non-managerial workers, contrary to human capital theory, decrease their voluntary absence only in the short-run. This short-term effect is

in line with results from a laboratory experiment on gift exchange by Gneezy (2004) who find a short-term effort increase of individuals when they are paid above market-clearing wages. Employees might hence positively reciprocate to training by a short-term decrease in absence hours. In addition, we find a reduced turnover probability for employees that have been trained in mainly general skills by the firm. This stands in contrast to the prediction of human capital theory as well and rather suggests that investments in general training may help firms to retain employees. The negative relation of training and turnover probability is further found to be larger for low-tenured employees. The effect may hence be strongest for the first trainings received within a firm. High-tenured employees may have already received several trainings, so that the marginal effect of training on turnover is reduced. Overall, the results suggest that training investments might lead to behavioral responses of employees.

Chapter 2

Social Ties and Subjective Performance Evaluations - An Empirical Investigation[1]

2.1 Introduction

In many jobs, not all aspects of employee performance are objectively measurable. Therefore, organizations frequently use subjective performance evaluations to assess the employees' contributions. Theoretical work in the economics and accounting literature has argued that the use of subjectivity in performance measures can strengthen incentive setting as more facets of the job can be appraised (Baker et al. (1994), Baiman and Rajan (1995)).[2] On the other hand the use of subjective components in evaluations raises issues of biased ratings which can cause substantial inefficiencies (see for example Prendergast and Topel (1993), Murphy and Cleveland (1995)). In a subjective assessment "*human judges other humans*" (Milkovich and Wigdor (1991)) which for instance may open the door to favoritism, so that supervisors can follow their personal social preferences and bias the outcome of the evaluation. A biased performance evaluation can, for instance, lead to

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

[2] Empirically, Gibbs et al. (2003) present survey evidence that subjectivity in bonus allocation complements weaknesses of objective performance measures.

an inefficient allocation of workers to tasks or jobs (Prendergast and Topel (1996)) or to a failure to identify training needs of employees when they are judged too leniently. Therefore, it is important to investigate potential distortions in subjective evaluations in a real organizational context which we do here and thus contribute to the progress of "*understanding how subjective assessments are made*" (Prendergast (1999): 57).

A key observation in the literature is that subjective performance evaluations tend to be lenient (Jawahar and Williams (1997), Moers (2005), Berger et al. (2010), Bol (2011)). Prendergast and Topel (1996) and Prendergast (2002) analyze subjective appraisals in economic models assuming that supervisors, while having some intrinsic preference for accurately reporting the true performance, also care for the welfare of their subordinates. This leads to a basic trade-off between accuracy and leniency and it directly results that evaluations are the more lenient, the stronger the supervisor's social preferences towards the evaluated subordinate. Based on this reasoning, we argue that a closer social proximity between supervisor and subordinate should lead to better performance ratings even when there are no differences in actual performance. The aim of this study is therefore to empirically investigate whether social ties between supervisor and subordinate can bias the outcome of subjective evaluations.

To our knowledge, the connection between the degree of acquaintance between rater and ratee and rating biases has only been analyzed in the psychological literature (see for instance Cardy and Dobbins (1986), Varma et al. (1996), or Lefkowitz (2000)). Most of these studies are either laboratory experiments with students or they lack objective measures of performance. For example, Kingstorm and Mainstone (1985) study the connection between personal acquaintance and task acquaintance (i.e. the level of the supervisor's familiarity with the employees tasks) on ratings of sales employees.

For our analysis, we use a 4-year panel data set from personnel records of a call center organization. A special feature is that the data set covers information on subjective performance evaluations and other objective performance measures. Incorporating the performance measures in the analysis helps us to discover systematic distortions in the evaluation process.

2.2 Measuring Social Proximity

The key hypothesis of our study is that – controlling for objective measures of performance – ratings are the higher the closer the social proximity between rater and ratee. Based on our data set, we use two proxies for social proximity. First, we suppose that the strength of the personal relationship between supervisor and subordinate depends on the size of the group evaluated. We analyze the effect of the number of employees a supervisor has to evaluate per year on the result of subjective evaluations and expect more lenient results for supervisors in smaller groups where the personal contact is closer. Second, we expect more lenient ratings for employees who have worked for the same supervisor a longer period of time. It is important to stress that our data set allows us to disentangle the different drivers of performance assessments, as there is a frequent reallocation of call center agents and supervisors between different teams and we can observe a number of more objective measures of performance.

An underlying assumption is of course that the frequency of interaction increases social proximity. There is quite substantial evidence backing this claim. In a very exhaustive psychological review on social proximity Baumeister and Leary (1995) for instance conclude that "*...several other studies suggest how little it takes (other than frequent contact) to create social attachment*". In an economic experiment Glaeser et al. (2000) show that the time since a first meeting between two interaction partners has a significant positive effect on the amount of money transferred in a trust game. Brandts and Solà (2010) study the effect of personal relations on distributive decisions and find discrimination against the subjects that are not personally known to the distributor.[3]

[3] Also some experimental studies started to invite subjects to the lab that have already known each other before (friends) and subjects that meet for the first time (strangers) to identify an effect of social ties. For example Abbink et al. (2006) investigate an effect of social ties in an experimental microfinance experiment. They find a more generous behaviour in repayment decisions between group members in a "friends"-treatment.

2.3 Institutional Background

We investigate personnel data of call center employees from an international company with headquarter in Germany. The data covers one German subsidiary between 2004 and 2007. The investigated subjects are call-agents whose main task is to deal with service queries over the telephone from clients who bought technical products. The business activities of the company are organized in departments, of which we observe a total of 12 in the full sample over the years. Departments with supportive and administrative activities like Human Resources, Accounting and IT are excluded because no objective performance measures as for the call center agents are observed. In addition, the performance review system differs for those departments. The company offers call center services to large business customers who outsource their technical support. Due to organizational and contractual changes in the client structure, not all departments exist over the five years: only two exist in the whole five years, three departments in four years, three in three years, four in two years and three departments only in one year. 11 of these departments are so-called "Inbound"-projects receiving calls from end costumers for a client, for instance a computer production firm, to answer technical or administrative queries.

A department consists of about 1 to 2 team leaders with leadership authority, one communication coach, one floor manager, several so-called second level and first level agents. The communication coach is responsible to train the communication skills of the agents while the floor manager is planning the service schedule and therefore controlling the capacities. Second level agents are promoted first level agents who, while still answering calls, also serve as a link between the team leader and the first level agents.

The subsidiary has implemented a subjective performance evaluation system demanding an overall evaluation of every agent by the team leader once a year according to different criteria. The results of the subjective evaluation do not affect monetary compensation directly but are important for instance for promotion decisions and the identification of training needs. The evaluation data is stored in an internal database with the exact time period the

evaluation is referring to. Employees that just entered the company or received a negative evaluation are forced to be rated again after six months. The supervisor can rate the employee for each criterion on a scale from 1 to 5, where 5 is the highest rate and 3 means "to be up to standard". Additionally every criterion is complemented by a behavioral statement. An important point is that the supervisor can access other performance measures which are stored in an internal database. These measures are collected on a monthly basis. The quality of the work is assessed by a so-called Transaction Monitoring (TM) tool. Calls are either followed by a second level agent sitting beside the monitored agent or recorded without the agent being informed. This randomly selected call is then evaluated according to a quite narrowly defined rating sheet and the test is passed when at least reaching $80 - 100\%$ of the maximal score. The speed of work is evaluated with the so-called Average Handling Time (AHT). It describes the average time an agent needs to process a call and can be broken down to hourly scores. A third objective performance measure are the days of absence during the subjective performance evaluation period (one year).

2.4 Data Set and Empirical Approach

At the end of the appraisal criterion catalogue the assessor is always asked to give an overall rating. We use this item as dependent variable throughout our analysis. The item is scaled on a 5- point likert-scale with values from 1 to 5 where 5 indicates the best value "far above requirements" and 1 indicates the lowest value "far below requirements".

We estimate the following specification:

$$Y_{it} = \alpha + \beta X_{it} + \vartheta V_{it} + \gamma I_{it} + \upsilon_t + a_i + \varepsilon_{it}$$

where Y_{it} is the individual rating of an agent i who is evaluated at time t. X_{it} represents the main indicators for social proximity which will be explained in the following and the vector V_{it} measures the objective performance measures for worker i in period t. I_{it} are further worker characteristics, υ_t year

dummies and a_i is an individual fixed effect. As the dependent variable is measured on an ordinal scale we additionally run ordered probit regressions. As using fixed effects in probit models may lead to inconsistent parameter estimates (Hsiao (2003) p.194) we rescaled the dependent variable as suggested by Van Praag and Ferrer-i Carbonel (2004) according to their probit-adapted OLS (or POLS) approach.[4] The idea of this approach is to create a new dependent variable by assigning to each ordinal category the z-value corresponding to the cumulative frequency of the category. This new independent variable is now cardinally measured and can be used in standard linear panel models. The rescaling makes the coefficients of the linear estimations similar to those of ordered probit estimations.

In our analysis, we apply two main indicators for social proximity. First, group size is measured by the quantity of evaluations an assessor conducted per year. As we keep only one evaluation per employee and year in our data set the number of evaluated employees per supervisor measures the group size. Hence, for every supervisor the absolute number of evaluations conducted per year is summed up in a variable called "assessments per year". Second, a dummy variable is introduced indicating that an appraisal has been conducted by the same supervisor the year before.

A typical problem of studying performance appraisal data is that distortions are hard to detect as the true performance is typically not observable to the researcher (see for instance the discussion in Kane et al. (1995)). Hence, it is hard to measure whether an employee received a good appraisal because of good performance or whether the appraisal was biased for instance due to favoritism or social preferences. A key feature of our data set is that besides the subjective evaluation we observe a number of more objective measures of performance. We can therefore control for objective measures of performance to exclude that the results are driven simply by differences in productivity. But more importantly, in the company we study, employees move between teams and supervisors quite frequently, which helps us to identify reasons for biased evaluations by using panel regressions.

Performance measures used as control variables are the average result of

[4]We thank an anonymous reviewer who suggested to follow this approach.

the Transaction Monitoring, the sum of the absence days during the period covered by the subjective performance evaluation, and two dummies measuring the Average Handling Time. These two dummies are generated as follows: One of the dummies indicates that the AHT value of an agent was below 90% of the mean AHT within his group in the considered year and the other one indicates that the AHT exceeded the mean value. The reason for this structure is that the company's objective is to make optimal use of capacity by having shorter calls but also to provide an acceptable quality. Other control variables cover individual-specific characteristics like age, age^2, tenure and sex and unit-specific attributes such as average age in the unit, or the percentage of women per unit. Additionally a dummy variable is included indicating whether a supervisor was conducting an appraisal for the first time in his or her career.[5]

We restrict our sample to full-time employees during the years 2004−2007. Additionally we only consider first level call center agents as there are different evaluation formats in use for different hierarchical levels. We dropped a few observations ($n = 22$) for which two evaluations have been stored in the data base for the same evaluation period. Since assigned values of the objective performance measures (that are partially measured on a daily basis) depend on the specific evaluation period we dropped the observations with missing details about the exact period, so that we reduced the sample to the observations complete in this respect. After these selection processes our sample consists of 520 employee-year observations. These agents are in total employed in 12 different departments and are evaluated by 18 different supervisors. The 520 observations cover 386 different individuals that have been assessed one to three times during the 4 years. There are very high turnover rates in the call center. Hence, only 33.7% of these individuals have been evaluated several times. Descriptive statistics of the main variables are presented in table 2.9 in the appendix. Please note that nearly 89% of the observations received a 3 ("fulfilled requirements") which affirms a "*managers'*

[5]Landy and Farr (1980), for instance, state that younger supervisors tend to evaluate more negatively than their more senior colleagues do. Hence, it is important to control for this effect.

tendency to assign uniform ratings to employees" (Murphy (1992)).

2.5 Results

We first look at the distribution of appraisal grades for small (less than 15 agents assessed by the supervisor per year), middle-sized (between 15 and 30 agents) and large groups (more than 30 agents) as shown in table 2.1. Indeed the table already indicates that better grades seem to be more frequent in smaller groups. The frequency of grade 4 is, for instance, twice as high in groups with less than 15 as compared to groups with more than 30 employees.

Grades Distribution (in %)	1	2	3	4	5
Small groups (< 15)	0	3.76	89.47	6.02	0.75
Middle-sized groups (≥ 15 & < 30)	0	9.13	86.31	4.18	0.38
Large groups (≥ 30)	0	5.63	91.34	3.03	0

Table 2.1: Distribution of Appraisal Grades by Group Size

As a starting point, simple pooled regressions regarding the effects of the number of assessed employees per assessor are shown in table 2.2 reporting robust standard errors clustered for teams. Column (1) shows the OLS regression without controlling for objective performance measures. The coefficient for the variable counting the number of assessments per supervisor-year is negative and significant at the 5%-level. In specification (2) the four objective performance measures are added. The coefficient of the assessments per year becomes stronger and achieves a significance level of 1%. Hence, in line with our hypothesis appraisals in smaller units are indeed more lenient. Ordered probit regressions confirm this result (columns (3) and (4) of table 2.2).

The coefficients of the objective performance measures show the expected signs. High Transaction Monitoring results positively affect the overall assessment, while the days of absence have significantly negative impact. The dummy variables for the AHT score boundaries have the expected sign but are insignificant. Having an assessor who has never rated before has also the

	Pooled OLS[a]		Pooled Ordered Probit	
Overall appraisal	(1)	(2)	(3)	(4)
Assessments per year	-0.007**	-0.010***	-0.017**	-0.026***
	(0.003)	(0.003)	(0.007)	(0.006)
TM		0.018***		0.047***
		(0.005)		(0.009)
Days of absence		-0.004***		-0.012***
		(0.001)		(0.003)
Over 100% AHT		-0.053		-0.155
		(0.052)		(0.152)
Under 90% AHT		0.020		0.041
		(0.073)		(0.176)
New Assessor	-0.429**	-0.513***	-0.888***	-1.142***
	(0.181)	(0.155)	(0.317)	(0.256)
Female	-0.139	-0.130	-0.335	-0.351
	(0.095)	(0.095)	(0.250)	(0.287)
Tenure	0.029***	0.037***	0.076***	0.116***
	(0.010)	(0.011)	(0.026)	(0.028)
Constant	3.269***	1.252		
	(1.100)	(0.834)		
Observations	520	520	520	520
R^2	0.098	0.159		
Pseudo Likelihood			-193.95661	-177.04632

[a]Dependent variable "overall appraisal" is cardinalized according to the approach by Van Praag and Ferrer-i Carbonel (2004). Robust Standard errors in parentheses. Clustered on team level. Control variables include age, age squared, year dummies, the share of women and average team age. ***p<0.01, **p<0.05, *p<0.1

Table 2.2: Number of Assessments: OLS and Ordered Probit

anticipated negative impact (significant on the 1%- level in columns (2), (3) and (4)) in the estimations.

While we consider it quite unlikely that team size is endogenous as it is mainly driven by client demands and we control for several measurable aspects of performance, our data allows us to go one step further and investigate panel data to control for further unobservable heterogeneity (such as individual abilities not captured by the objective performance measures). The results of fixed and random-effects regressions are reported in table 2.3 and confirm the previous observations in all specifications. The model predicts that a specific employee switching from a smaller to a larger group will receive an inferior evaluation even if his true performance is unaffected.[6] A Hausman-Test does not reject that the a_i are uncorrelated with the explanatory variables (p=0.8829). Still, the fixed effects model is our most preferred specification as we have a complete data set of all call center agents in the given unit.

Next we analyze the effect of a repeated assessment by the same supervisor on performance evaluations. We therefore created a dummy variable indicating whether the employee has been evaluated by the same assessor before.

Table 2.4 shows the distribution of grades dependent on whether there has been a previous assessment by the same supervisor. Note that 5.08% of those employees who have been assessed by the same supervisor before receive a good grade of 4 while only 2.88% of those who had been appraised by a different supervisor before receive this grade. Furthermore, supervisors who rate an employee for the first time give the low grade 2 more than five times as often as supervisors who have evaluated the same employee before. This could indicate that more social proximity leads to better grades.

[6] Please note that in the fixed effects regressions, we omitted the linear terms of tenure and age because the effects cannot be distinguished from the time effects when controlling for year dummies.

	Random effects - OLS[a]		Fixed effects - OLS[a]	
Overall appraisal	(1)	(2)	(3)	(4)
Assessments per year	-0.007***	-0.010***	-0.015***	-0.016***
	(0.002)	(0.002)	(0.005)	(0.005)
TM		0.019***		0.014*
		(0.004)		(0.009)
Days of absence		-0.004***		-0.003
		(0.002)		(0.004)
Over 100% AHT		-0.042		0.131
		(0.064)		(0.130)
Under 90% AHT		0.019		-0.010
		(0.068)		(0.143)
New Assessor	-0.436***	-0.522***	-0.651**	-0.763**
	(0.140)	(0.133)	(0.313)	(0.321)
Female	-0.143**	-0.134*	-	-
	(0.071)	(0.070)		
Tenure	0.029	0.037**	-	-
	(0.018)	(0.019)		
Constant	3.344***	1.320	7.045***	5.878*
	(0.885)	(0.930)	(2.714)	(3.055)
Observations	520	520	520	520
Overall R^2	0.097	0.159	0.126	0.159

[a]The dependent variable is cardinalized (Van Praag and Ferrer-i Carbonel (2004)).
Robust Standard errors in parentheses. Further control variables include age, age squared, year dummies, the share of women and average team age.
In the FE regressions, the linear age and tenure terms are excluded.
***p<0.01, **p<0.05, *p<0.1.

Table 2.3: Number of Assessments: Random and Fixed Effects

Grades Distribution (in %)	1	2	3	4	5
Different supervisor	0	8.65	88.46	2.88	0
Same supervisor	0	1.67	93.22	5.08	0

Note: Only repeated appraisals taken into account.

Table 2.4: Distribution of Appraisal Grades by "Repeated Assessment"

	New assessment	Repeated assessment
% of evaluated employees with lower grade in t compared to $t-1$	14.58	5.36
% of evaluated employees with same grade in t compared to $t-1$	80.21	83.93
% of evaluated employees with higher grade in t compared to $t-1$	5.21	10.71

Note: Only repeated appraisals taken into account.

Table 2.5: Changes of Grade from t-1 to t in Dependence of Repeated Assessment by the Same Supervisor

It is also interesting to compare changes in grades for given employees (table 2.5): When being appraised by the same supervisor a grade improvement occurs twice as often as when the supervisor has changed (10.71% in comparison to 5.21%). On the other hand, the probability that an employee gets a worse grade is three times as high in case of an assessment by a different supervisor (14.58% in comparison to 5.36%). Hence, for a given employee the chances to obtain a better grade are higher if he is repeatedly evaluated by the same supervisor.

Of course, the repeated assessment dummy may capture also simple experience effects. Hence, it is very important to control for firm tenure. The results of OLS and ordered probit regressions are reported in table 2.6. Columns (1) and (4) contain the results for specifications without further performance measures while we control for these measures in specifications (2) and (5). We find that employees receive a better grade when they are repeatedly assessed by the same supervisor as compared to employees of the same tenure attaining the same performance measure values who are assessed by a different supervisor.

	Pooled OLS[a]			Pooled Ordered Probit		
Overall appraisal	(1)	(2)	(3)	(4)	(5)	(6)
Repeated Appraisal Same Supervisor	0.150*	0.145*	0.103	0.364*	0.366**	0.284*
	(0.079)	(0.074)	(0.065)	(0.187)	(0.175)	(0.172)
TM		0.012*	0.013*		0.033**	0.035**
		(0.007)	(0.007)		(0.014)	(0.015)
Days of absence		-0.005***	-0.004***		-0.011***	-0.011***
		(0.001)	(0.001)		(0.003)	(0.003)
Over 100% AHT		-0.060	-0.063		-0.146	-0.165
		(0.050)	(0.050)		(0.140)	(0.141)
Under 90% AHT		0.003	0.003		0.007	-0.011
		(0.067)	(0.067)		(0.164)	(0.162)
Female	-0.162	-0.156	-0.148	-0.395	-0.418	-0.413
	(0.104)	(0.108)	(0.107)	(0.257)	(0.291)	(0.301)
Tenure	0.035**	0.046***	0.044***	0.086***	0.127***	0.125***
	(0.015)	(0.014)	(0.012)	(0.031)	(0.029)	(0.028)
New Assessor			-0.333*			-0.673**
			(0.179)			(0.311)
Constant	2.102	0.206	-0.299	-6.036**	-1.423	-0.0568
	(1.386)	(1.506)	(1.223)	(2.655)	(3.302)	(2.710)
Observations	520	520	520	520	520	520
R^2	0.068	0.109	0.126			
Pseudo Likelihood				-200.755	-189.130	-185.867

[a]The dependent variable is cardinalized according to Van Praag and Ferrer-i Carbonel (2004). Robust standard errors in parentheses. Clustered on team level. Control variables include age,age squared, year dummies, the share of women and average team age. ***p<0.01, **p<0.05, *p<0.1

Table 2.6: Repeated Appraisals: OLS and Ordered Probit

The two further specifications (3) and (6) additionally include a "new assessor"-dummy indicating that a supervisor had no prior experience with evaluations. Note that this reduces the effect size for the repeated appraisal. While the effect of repeated appraisals becomes insignificant in the OLS regressions it stays weakly significant in the ordered probit regression. Hence at least part of the effect is driven by the tendency of inexperienced supervisors to assign worse grades. But again, it seems very important here to control for unobserved heterogeneity. To see that, note that the comparison of the results with and without the objective performance measures shows an increase of the tenure coefficient in columns (2) and (5). Due to on the job human capital formation we would usually expect a better performance of employees with higher tenure and hence a decreasing tenure coefficient when objective performance measures are included. Interestingly, we observe the opposite pattern as the tenure coefficient gets even stronger. This can be best understood when considering the two graphics in figure 2.1 and 2.2 which illustrate average Transaction Monitoring scores and days of absence per year of tenure. The TM results do not increase with tenure and even fall beginning with the fifth year of tenure and the days of absence consistently increase in the data set. These developments have two different reasons. First of all, the jobs in the call center are typically regarded as stressful, hence absence rates increase and performance seems to go down. In addition, there are selection effects as able first level agents will be promoted to the second level and poorly performing agents leave the company.

To control for unobserved heterogeneity and selection effects we therefore again ran random and fixed effects regressions (see table 2.7). Again, the Hausman Test does not reject that the a_i are uncorrelated with the explanatory variables (p=0.6883).

	Random Effects - OLS[a]			Fixed Effects - OLS[a]		
Overall appraisal	(1)	(2)	(3)	(4)	(5)	(6)
Repeated Appraisal Same Supervisor	0.156**	0.152*	0.105	0.267*	0.351**	0.309**
	(0.077)	(0.078)	(0.077)	(0.143)	(0.143)	(0.136)
TM		0.013***	0.014***		0.017*	0.019*
		(0.004)	(0.004)		(0.010)	(0.010)
Days of absence		-0.005***	-0.004***		-0.001	-0.001
		(0.002)	(0.002)		(0.004)	(0.004)
Over 100% AHT		-0.048	-0.052		0.074	0.077
		(0.066)	(0.065)		(0.133)	(0.131)
Under 90% AHT		0.002	0.001		-0.029	-0.034
		(0.070)	(0.069)		(0.148)	(0.145)
Female	-0.166**	-0.161**	-0.154**	-	-	-
	(0.075)	(0.074)	(0.073)			
Tenure	0.034*	0.046**	0.044**	-	-	-
	(0.019)	(0.019)	(0.018)			
New Assessor			-0.327**			-0.196
			(0.130)			(0.253)
Constant	2.068***	0.205	-0.293	2.141	3.364	3.550
	(0.798)	(0.978)	(0.961)	(2.162)	(3.242)	(3.233)
Observations	520	520	520	520	520	520
Overall R^2	0.068	0.109	0.126	0.055	0.122	0.127

[a]The dependent variable is cardinalized according to Van Praag and Ferrer-i Carbonel (2004). Robust standard errors in parentheses. Clustered on team level. Control variables include age, age squared, year dummies, the share of women and average team age. In the FE regressions, the linear age and tenure terms are excluded. ***p<0.01, **p<0.05,*p<0.1

Table 2.7: Repeated Appraisals: Random and Fixed Effects

The repeated appraisal dummy is again significantly positive in all fixed effects specifications. Hence, a given employee at a given point in time indeed obtains better grades when he is evaluated by a supervisor he is familiar with as compared to a situation in which he is evaluated by a different supervisor.

Note that it could be argued that supervisors who have evaluated the same person before, can more accurately appraise the work of the employee as they are able to observe them over a longer time. However, while this may lead to more differentiated grades it should not lead to grades which are better on average such as we observed. Moreover, as shown in table 2.10, the standard deviation of assessments by the same supervisor is smaller rather than larger which also makes such a mechanism implausible.

Finally, we estimate regressions in which we use both proxies for social proximity in the same specification. The results are similar when we include both proxies for social ties, the unit size and the dummy for the repeated appraisal by the same supervisor as is shown in table 2.8. But here the effects of team size are somewhat more robust than those of repeated appraisals. Again, the Hausman Test of the first specification revealed that there is no systematic difference between the random and the fixed effects regression (p=0.5742). In a final robustness check, we included assessor fixed effects to the OLS baseline specifications because the personal characteristics of supervisors might be relevant for the subjectively given grades. As can be seen in table 2.11 in the appendix, sign and size of the coefficients stay robust when including assessor fixed effects but the effect of number of assessments is no longer significant. This is probably due to the small within-supervisor variation (i.e. team size did not vary sufficiently for given supervisors). However, the effect is highly significant for the repeated assessment (here there is natural within-supervisor variation) which confirms that for a given assessor, a repeated employee-supervisor interaction positively biases the performance evaluation.

	Random Effects - OLS[a]		Fixed Effects - OLS[a]	
Overall appraisal	(1)	(2)	(3)	(4)
Assessments per year	-0.007***	-0.010***	-0.007*	-0.014***
	(0.002)	(0.002)	(0.003)	(0.005)
Repeated Appraisal Same Supervisor	0.144*	0.068	0.347**	0.209
	(0.080)	(0.079)	(0.144)	(0.140)
TM	0.015***	0.018***	0.014	0.015*
	(0.004)	(0.004)	(0.010)	(0.009)
Days of absence	-0.005***	-0.004***	-0.002	-0.002
	(0.002)	(0.002)	(0.004)	(0.004)
Over 100% AHT	-0.036	-0.035	0.103	0.146
	(0.065)	(0.064)	(0.135)	(0.133)
Under 90% AHT	0.016	0.022	-0.014	-0.011
	(0.069)	(0.068)	(0.147)	(0.142)
Female	-0.152**	-0.137*	-	-
	(0.072)	(0.070)		
Tenure	0.041**	0.036*	-	-
	(0.019)	(0.019)		
New Assessor		-0.508***		-0.628*
		(0.134)		(0.332)
Constant	1.479	1.357	4.043	5.430*
	(0.953)	(0.936)	(3.134)	(3.058)
Observations	520	520	520	520
Overall R^2	0.126	0.160	0.136	0.170

[a]The dependent variable is cardinalized according to Van Praag and Ferrer-i Carbonel (2004). Robust standard errors in parentheses. Clustered on team level. Control variables include age,age squared, year dummies, the share of women and average team age. In the FE regressions, the linear age and tenure terms are excluded. ***p<0.01, **p<0.05, *p<0.1

Table 2.8: Number of Assessments and Repeated Appraisals: Random and Fixed Effects Estimations

2.6 Conclusion

We investigated possible distortions in subjective performance appraisals and found evidence for the hypothesis that subjective performance is biased when there is a closer social proximity between supervisor and subordinates. Our analysis shows that the size of the work unit has a negative impact on grades in subjective performance evaluations. Controlling for objective performance measures employees in large units received worse evaluations than employees in smaller units. We also observed that employees who have been evaluated by the same supervisor before receive better ratings. Both results also hold in fixed and random effects regressions, where we thus controlled for unobserved heterogeneity in abilities. Hence, we conclude that a person with given experience and performance receives lower ratings when moving to a larger team or when getting a new supervisor.

It is important to note that such distortions in subjective evaluations may have substantial consequences for performance as, for instance, rating leniency may negatively affect differentiation in performance grades, at least for those employees who are well acquainted with their supervisor. There is now a number of studies investigating the performance consequences of differentiated ratings. Engellandt and Riphahn (2011), for instance recently have analyzed a large data set from a multinational company finding that effort as measured by overtime hours is higher in departments in which individual performance evaluations are more flexible over time. Bol (2011) reports a positive effect of leniency bias on performance improvement while she finds a negative impact of centrality bias on performance. Berger et al. (2010) have conducted a laboratory experiment observing that lenient ratings led to a subsequent drop in performance.

Moreover, systematic distortions in appraisals may have a negative effect on employee morale. When appraisals have an impact on subsequent wage increases, distortions may reduce the connection between performance and wage increases and this may have detrimental effects on performance.

Hence, it should be worthwhile for firms to invest in avoiding potential biases. Firms may, for instance, consider training supervisors to become

aware of potential distortions in their evaluation behavior. Moreover, it seems useful to confront supervisors with systematic deviations between objective performance measures and their subjective assessments.

Finally, our results also indicate that firms must be cautious when using subjective performance evaluations to compare employees across departments or to assign bonuses. There is a bias in favor of employees from smaller groups and employees who have been acquainted with the supervisor for longer periods of time. These effects have to be taken into account when decisions on promotions or layoffs are made forcing a firm to rank employees across departments.

2.7 Appendix of Chapter 2

Variable Group and Description	Mean	SD
Dependent Variable		
Overall assessment	2.967	0.336
Indicators for social ties		
Assessments per year (by supervisors)	32.994	21.025
Repeated Appraisal Same Supervisor (Dummy)	0.113	0.317
Objective Performance Measures		
Result Transaction Monitoring (TM)	90.554	8.992
Over 100% AHT per group-year (Dummy)	0.462	0.499
Under 90% of mean AHT per group-year (Dummy)	0.285	0.452
Days of absence	13.611	18.804
Individual Characteristics		
Tenure	2.754	1.988
Dummy female $(1/0)$	0.387	0.487
Age	32.323	9.260
$(\text{Age})^2$	1130.36	661.311
Characteristics of assessor/ assessor unit		
Average Age of unit	31.957	1.709
Share of female employees	0.372	0.197
Dummy new assessor $(1/0)$	0.077	0.267

Note: The table describes all main variables on the basis of N=520 observations.

Table 2.9: Descriptive Statistics

Grades by new assessments	Mean	Sd
Different supervisor	2.975	0.356
Same supervisor	3.033	0.258

Table 2.10: Mean and Standard Deviation by "Repeated Assessment"

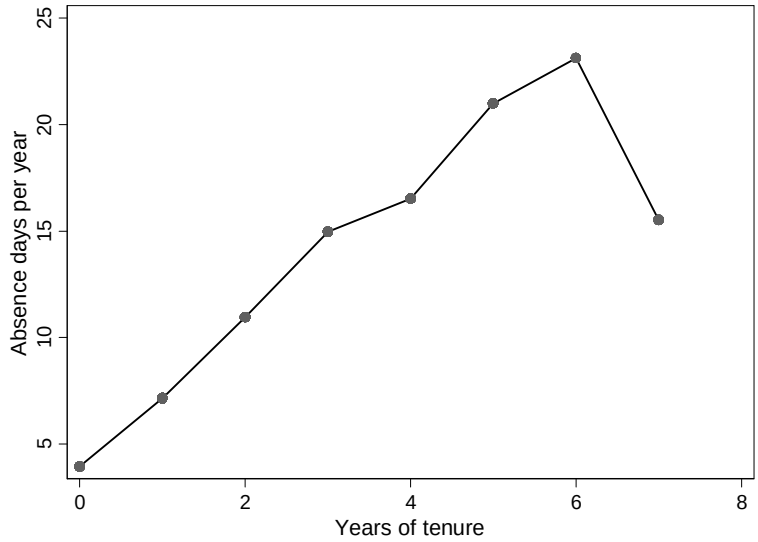

Figure 2.1: Absence Days by Years of Tenure

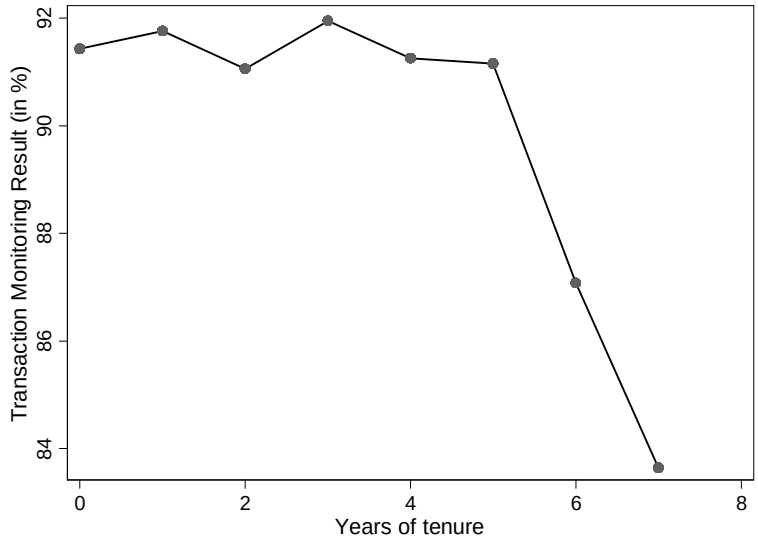

Figure 2.2: Transaction Monitoring by Years of Tenure

	OLS with assessor fixed effects[a]		
Overall appraisal	(1)	(2)	(3)
Assessments per year	-0.004		-0.003
	(0.003)		(0.003)
Repeated Appraisal Same Supervisor		0.226***	0.222***
		(0.083)	(0.082)
TM	0.019***	0.019***	0.019***
	(0.005)	(0.005)	(0.005)
Days of absence	-0.003*	-0.003*	-0.003**
	(0.002)	(0.002)	(0.002)
Over 100% AHT	-0.095	-0.082	-0.079
	(0.063)	(0.062)	(0.062)
Under 90% AHT	-0.032	-0.033	-0.029
	(0.069)	(0.069)	(0.069)
Female	-0.118*	-0.126*	-0.123*
	(0.065)	(0.065)	(0.065)
Tenure	0.027	0.021	0.020
	(0.018)	(0.017)	(0.017)
New Assessor	-0.262	-0.139	-0.227
	(0.276)	(0.247)	(0.272)
Assessor Fixed Effects included	yes	yes	yes
Constant	-6.913	-7.007	-6.522
	(5.351)	(5.293)	(5.354)
Observations	520	520	520
Overall R^2	0.233	0.240	0.241

[a]The dependent variable is cardinalized according to Van Praag and Ferrer-i Carbonel (2004). Standard errors in parentheses are clustered on individual level. ***p<0.01, **p<0.05, *p<0.1

Table 2.11: OLS Baseline Specification Including Assessor Fixed Effects

Chapter 3

Are Women Better Leaders? An Empirical Investigation of Gender Differences in Manager's Evaluation Behavior[1]

3.1 Introduction

As women remain rare as elite leaders and top executives in firms (Catalyst (2010)), many reasons for this observation are discussed in the economic, psychological, and management literature. Besides theories of taste-based or statistical discrimination (Becker (1957), Lazear and Rosen (1990), Booth et al. (2003), Bjerk (2008)), differences between men and women in general preferences towards a career or promotion tournaments (Konrad et al. (2000), Croson and Gneezy (2009), Niederle and Vesterlund (2007)) have been discussed. However, a further reason may be a gender difference in the ability of being a good leader. In economics, studies on leadership differences between men and women are scarce. The overall effect of women in corporate

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

boards on firm performance has been investigated (Carter (2003), Erhardt et al. (2003), Adams and Ferreira (2009)), but evidence is mixed.

In this study, we aim to contribute to the question of whether men and women differ with regard to their leadership qualities by investigating differences in their performance evaluation behavior. Managers who supervise employees in white collar jobs typically have to evaluate the performance of their employees in order to set incentives, decide about promotion, training needs or to allocate tasks when no objective performance measures are available (Prendergast (1999)). These performance evaluations are a core task when supervising other employees (Murphy and Cleveland (1995)), so that it is important to understand how men and women may differ with regard to this task. Specifically, we investigate whether female managers give more differentiated performance evaluations than their male colleagues.

A higher differentiation between employees is important especially when evaluations are tied to monetary incentives (see for example Prendergast (1999)). When bonus payments are tied to ratings, a higher differentiation implies an increasing marginal return to effort for employees (see for example Holmström (1979)). Employees anticipate that with more differentiated ratings, high effort is more likely to be rewarded while low effort is more likely to be punished which increases the incentive to exert greater effort. Moreover, differentiation helps in taking the right personnel decisions. Like Jack Welsh formulated (Welch (2003): 195): To improve workforce performance, an organization should be capable of "sorting out the A, B and C players" to promote A and to get rid of C players.

However, it has been claimed that supervisors in practice often compress the subjective performance evaluations relative to the true performance distribution of employees.[2] Typically, this bias is referred to as centrality bias. Compressed ratings imply a reduction of incentives due to the lack of rewards and punishments.

Indeed, some field and experimental studies have already confirmed the

[2]In their case study of Merck, Murphy (1992) give anecdotal evidence. For an overview of rating biases see Murphy and Cleveland (1995), Milkovich and Wigdor (1991) and Prendergast and Topel (1996), Prendergast and Topel (1993) and MacLeod (2003) for an economic model.

performance-enhancing effect of differentiated evaluations or bonus payments. Bol (2011) shows that a compression in ratings leads to a decrease in effort by high and low performers in a Dutch financial service company. Engellandt and Riphahn (2011) support this finding, showing that a higher variability in ratings increases individual performance. In an experimental study, Berger et al. (2010) investigate the forced-distribution-system as a measure for obliging supervisors to differentiate more in ratings, and they find that individual effort increases when supervisors are forced to differentiate more. Furthermore, Kampkoetter and Sliwka (2010) analyze the effect of bonus differentiation in the financial service industry confirming that higher differentiation in bonus payments increases individual performance. This evidence suggests firms to employ supervisors that sufficiently differentiate between their employees.

Besides assigning more differentiated grades to subordinates, a higher differentiation also includes the assignment of poor grades to low performing subordinates. This can be perceived as the difficult part of conducting performance appraisals in organizations as individuals are typically reluctant to assign poor grades in order to avoid negative consequences such as a violated supervisor-subordinate relationship or negative reactions by the employee (Harris (1994), Shore and Tashchian (2002), Yariv (2006)). The fear of negative consequences may lead to the overall tendency of being lenient in evaluations (Napier and Latham (1986), Shore and Tashchian (2002), MacLeod (2003)) leading to an increased compression of ratings. Jack Welch illustrated the difficulty of GE managers to assign low grades to their subordinates. He states that managers "will play every game in the book to avoid identifying their bottom 10" (Welch (2003): 198).[3] But not identifying the low performers may weaken the firm's productivity. Thus, assigning low grades is a crucial part of performance evaluations in organizations and so we also analyze whether women or men are more likely to assign poor grades to their subordinates.

[3]One example he gives is that a GE manager put a name of a man in the bottom 10 category who has died two month before, just to avoid the nomination of one team member.

A wide range of studies in the psychological literature have analyzed performance ratings with focus on information processing, rating accuracy, reactions to ratings, rater training and on rating formats (see Bretz et al. (1992) for extensive reviews of the literature). Further research has addressed the effect of social context on evaluation outcomes (see Levy and Williams (2004) for an overview) including personal supervisor characteristics. Only some studies looked at the impact of supervisor gender on performance evaluations so far (Benedict and Levine (1988), Furnham and Stringfield (2001) and Varma and Stroh (2001)), but ignore the analysis of differentiation. Two of the studies especially focus on evaluations in same-sex supervisor-subordinate dyads and find, based on experimental data sets, that performance ratings are higher when supervisors evaluate subordinates of the same gender (Furnham and Stringfield (2001), Varma and Stroh (2001)).

Our empirical analysis is based on a dataset of a large company headquartered in Germany for the years $2006 - 2008$. We observe the performance evaluations of every managerial worker in the firm and many other personal and job-specific characteristics. The data set is unique, as it allows us to identify work unit and the respective supervisor so that we can incorporate work unit, supervisor, and individual characteristics in our regression analysis. We use the standard deviation in grades and the share of poor grades for a workgroup as dependent variables. Due to the panel structure of the dataset, we are able to conduct fixed effects regressions to eliminate unobserved heterogeneity of a workgroup. Our analysis shows that women differentiate more between their employees and give lower grades more often than men. In this respect they are therefore rather better leaders on average as compared to men.

The paper proceeds as follows. Section 3.2 provides a discussion on why differences in the evaluation behavior between men and women can be expected due differences in personality, risk attitudes or social preferences and sums up the relevant literature. Section 3.3 introduces the data set and section 3.4 presents some descriptive results. Subsequently, section 3.5 to 3.7 report the results. Section 3.8 concludes.

3.2 Reasons for Gender Differences in Evaluation Behavior

Much economic research has been conducted to analyze gender differences in risk attitudes and social preferences (see, e.g. Holt and Laury (2002), Dohmen et al. (2006) and Croson and Gneezy (2009) for an overview) that may give explanations for gender differences in evaluation behavior. In the following, we aim at developing these arguments. A first robust insight from these studies is that women are more risk-averse than men (see Croson and Gneezy (2009) for a summary). Risk attitudes may be relevant in the differentiation between employees in two ways. First, higher risk aversion possibly implies less compression in evaluations. Given that compressed ratings may have negative incentive effects, organizations may want to monitor supervisors (Prendergast and Topel (1996)). If a central tendency of supervisors is thereby detected, supervisors may be fined. With a positive probability of bias detection, more risk averse individuals may hence shy away from taking that risk and evaluate in a more differentiated and accurate manner. Second, risk aversion can also cause more centrality bias in evaluations, as low grades for subordinates may cause a risk of negative reactions by subordinates (Napier and Latham (1986)).[4] Supervisors with higher risk aversion may want to avoid this risk and bias the evaluations toward the top of the scale, which leads to more compressed ratings. However, although there is some evidence that risk attitudes do not differ between men and women in the managerial population (Johnson and Powell (1994)), a recent study by Berger (2011) concludes based on the analysis of two data sets, that women are more risk-averse on every level of a hierarchy.

Second, experiments have shown that women more reciprocal than men (Eckel and Grossman (1996), see Croson and Gneezy (2009) for a review). Eckel and Grossman (1996) show in their paper that women are more likely

[4] MacLeod (2003) for example modeled optimal contracts with subjective performance evaluations and showed that if the agent´s and principal´s signals about the agent´s performance are weakly correlated, this leads to bias in evaluations through the avoidance of agency costs. Her furthermore shows that this bias reduced the incentive setting and leads to worse worker morale.

to reward fair behavior and punish unfair behavior. Hence, women may be more likely to reward employees with good grades when high effort is exerted and also to punish subordinates for low performance. This would predict that women differentiate more between their subordinates than men.

Third, women seem to be less likely to lie than men (Dreber and Johannesson (2008), Conrads et al. (2011)). Following the study of Gneezy (2005) examining deception, Dreber and Johannesson (2008) used a sender-receiver game in which the sender has a monetary incentive to send a deceptive message to the receiver. They find that men are much more likely to lie than women in order to secure a monetary benefit (Dreber and Johannesson (2008)). Conrads et al. (2011) confirm the lower lying inclination for women in their study on lying under piece-rate or team incentive scheme. Women may hence evaluate more truthfully and differentiate more between their subordinates.

Fourth, Croson and Gneezy (2009) argue that women are more inequality averse in their giving, which has been analyzed in dictator games. In performance ratings, more inequality aversion would imply more compressed evaluations. However, when considering the invested effort of employees, a higher inequality aversion of women may also imply a higher differentiation in ratings conditional on different investments of employees.

Furthermore, differences in personality traits between men and women can contribute to differences in evaluation behavior. Personality characteristics are generally studied based on the Big Five personality traits that are supposed to broadly cover the personality of a person (Costa and McCrae (1992)). Bernadin et al. (2000) who investigate personality traits as predictors of rating leniency, found that conscientiousness seems to be negatively related to leniency in evaluations while agreeableness was positively correlated with rater leniency. Similar correlations of personality traits can be expected with regard to the differentiation in grades. In psychological studies, women rated higher in neuroticism, agreeableness, extraversion and conscientiousness than men, which seem to be consistent across cultures (Schmitt et al. (2008)). Conscientious people are conceptually defined as organized, having the tendency to follow norms and rules, striving for excellence and

to think before acting (John and Srivastava (1999)). Managers with higher conscientiousness can be expected to evaluate in an accurate and thorough manner implying a higher differentiation in grades. Moreover, conscientiousness can be assumed to be negatively correlated with lying. On the contrary, agreeableness implies a prosocial and communal orientation towards others, causing agreeable people to be altruistic, trustful and modest (John and Srivastava (1999)). Managers who are more agreeable might therefore evaluate in a modest manner, leading to more compressed evaluations.

3.3 The Data Set

To analyze the differentiation in performance evaluations, we use personnel records of a large multinational firm based in Germany for the years $2006 - 2008$.[5] The data contains information about the subjective performance ratings of all employees in Germany. Because of confidentality reasons, personnel data of managers working at top two hierarchical levels had to be excluded. At the end of every year, supervisors in this firm have to evaluate their direct subordinates based on a predefined form. The outcome of this evaluation is a grade for *job performance* that ranges from 1 to 5 with 1 standing for "problematic", 2 for "partially meets expectations", 3 for "fully meets expectations", 4 for "exceeds expectations" and 5 for "outstanding".[6] In general, the grade determines a substantial part of the bonus payment an employee receives in the respective year.[7] Moreover, the policy of the performance management system was identical in the three observed years. By considering those employees as one unit who are working under the same

[5] We additionally observe the year 2009, but leave it out of the analysis because the performance management system was changed in that year.

[6] Please note that in the original scale of the firm, 1 denoted the best and 5 the worst grade. We have recoded the grade for interpretation reasons. In addition, there is no standard which is set for managerial employees individually. The performance standard that has to be attained for receiving a certain grade should hence not differ between employees.

[7] Please note that the ratio of bonus to total salary is increasing with the hierarchical level of managers in this firm. For the lowest managerial level, the ratio is 15% on average and it goes up to about 44% for the top managerial levels.

supervisor, we can look at the grades distribution per unit. In the following, we will refer to this level of aggregation as 'work unit'. Besides the information on performance grades assigned by the direct supervisor, the data set contains information on demographic characteristics of employees (length of service, age, gender) and job characteristics (hierarchical level, subgroup, fixed salary). We are further able to match supervisor characteristics (e.g. supervisor gender) and unit characteristics (e.g. female share in work unit) to the work unit. The data set set contains information on $5,775$ managerial employees which are observed several times over the years yielding $13,116$ employee-year observations[8]. By means of the supervisor ID, the employee-year observations can be aggregated to $3,392$ work unit-year observations.

Overall, the female share in the data is 18.0 percent, while 9.8 percent of the identified unit-supervisors who are evaluating employees are women. The overall share of women increased over the years from 17.2 percent in 2006 to 19.3 percent in 2008.

3.4 Descriptive Statistics on Performance Evaluations in the Firm

We begin our analysis by looking at the distribution of performance ratings in this company. Figure 3.1 shows a histogram of the pooled performance ratings over the years $2006 - 2008$.[9] Obviously, there is a central tendency in evaluations. About 71.6 percent of employees received a 3 for "fully meets expectations." Furthermore, the distribution is slightly skewed to the right with a higher share of employees having received a grade better than 3 than a grade worse than 3. Over the years $2006 - 2008$, only about 3.09 percent of the observations are evaluations with the lowest grades of 1 and 2. And similar to the study by Medoff and Abraham (1980) (p.709) and to the case

[8] We originally observe $15,152$ performance evaluations of employees over the three years but due to missing information on the supervisor ID, we can only use 87% of these for our analysis. One reason for this missing information is the exclusion of the top level employees in the analysis.

[9] Please note that the distribution did hardly change over the years (see figure 3.7 in the appendix).

study by Murphy (1992) (p. 40), about 96 percent of evaluations are crowded into two categories in our data set .[10] The pooled average rating is 3.23 with a standard deviation of 0.51, and mean values and standard deviation of ratings hardly change over the years (in 2008, the mean (standard deviation) of grades is 3.21 (0.51) while it is 3.26 (0.53) in 2007 and 3.21 (0.48) in 2006).

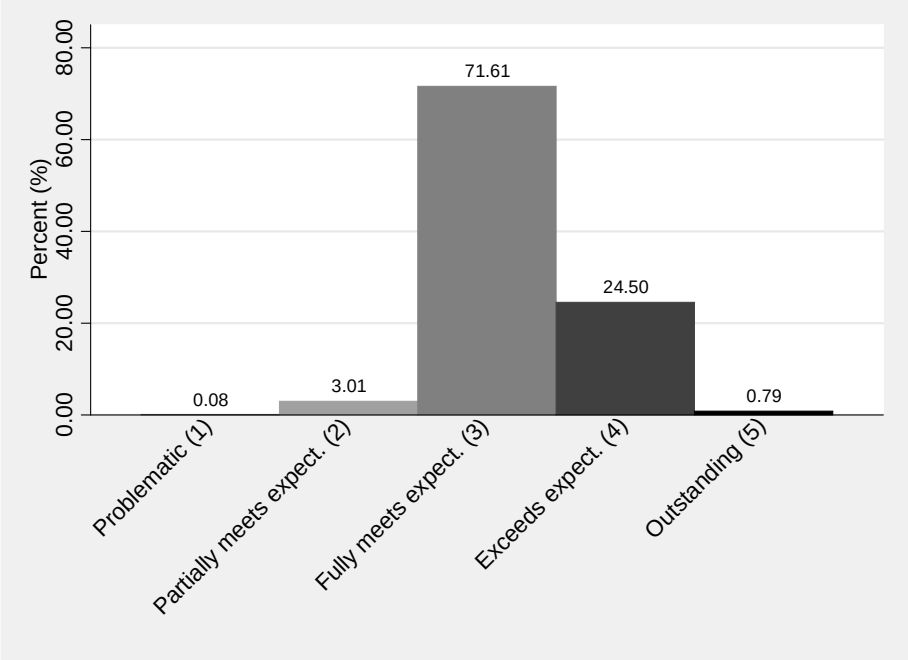

Figure 3.1: Distribution of Grades for the Years 2006-2008 (pooled)

As we aim to investigate the differences in evaluations by gender of the supervisor, we separated the distribution of performance evaluation with respect to the supervisor gender in figure 3.2. About 9 percent (1169 out of 13116 employee evaluations) are conducted by female managers (see table 3.8 in the appendix). Although the means in performance grades given by male and female supervisors do not differ (see table 3.8 in the appendix), figure 3.2 shows that on average female supervisors differentiate more in their ratings

[10]Medoff and Abraham (1980) report that in both analyzed firms, about 95% of evaluations are concentrated in two categories. In the case study of Merck (Murphy (1992)), even 96% of evaluations fall into two categories, but categories are are subdivided in three subcategories.

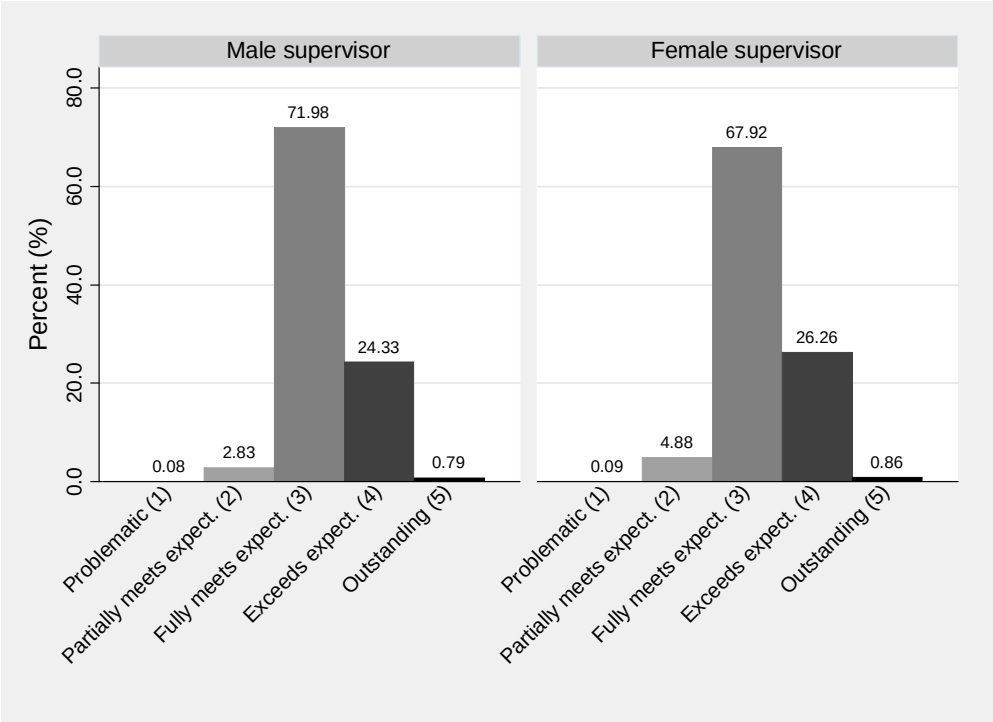

Figure 3.2: Distribution of Grades for Male and Female Supervisors (2006-2008)

than male supervisors do. Women rate a lower share of their employees with the middle grade of 3 (t-test, p=0.0033). And the difference in the share of giving a 2, which equals "Partially meets expectations," is substantial: Female managers give a 2 nearly twice as often as men (4.9% compared to 2.8%) which is significant on on the 1% level (t-test, p=0.0001).[11] Hence, women seem to be less reluctant to assigning a grade that clearly indicates poor performance. Identifying low performers is the inconvenient part of performance evaluations (Harris (1994), Shore and Tashchian (2002), Yariv (2006)) and illustrated by Grote (2005) who says (p. 27):"... too many managers display a reluctance to make meaningful differentiations among the troops, preferring to live in a Lake Wobegon coccon where all the children

[11]To put it in absolute numbers: Women assign the grade 2 to 57 out of 1112 of their employees while men assign the same grade to 338 out of 11609 employees.

are above average." The result is confirmed by figure 3.3, which shows that female supervisors gave poor grades more frequently than men in every year. The difference in the share of poor grades is statistically significant in 2006 and 2008 ($p = 0.0131$ in 2006 and $p = 0.0026$ in 2008).

Bonus Difference (€)	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Grade 2 - Grade 3	3,216	4,891	5,413	10,794	11,864	58,079
Grade 3 - Grade 4	2,613	2,933	3,247	5,988	11,017	13,452
Relative Loss						
Grade 2 - Grade 3	4.51%	5.10%	4.73%	7.94%	7.58%	32.14%
Grade 3 - Grade 4	3.59%	3.08%	2.89%	4.42%	7.03%	7.44%

Table 3.1: Marginal Return of Bonus Payments for Different Grades

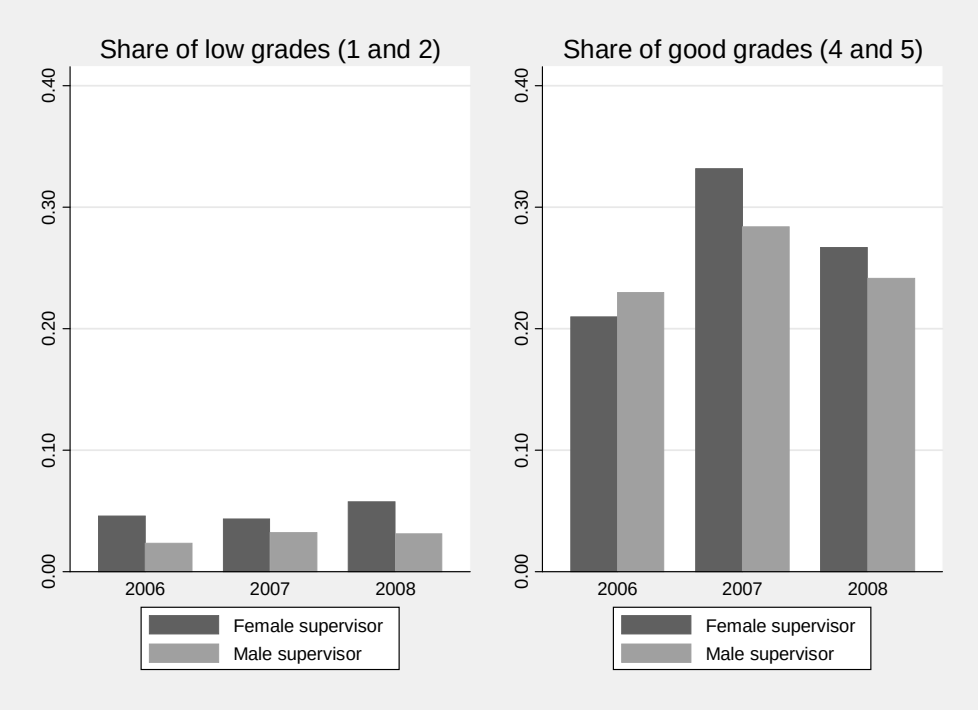

Figure 3.3: Share of High and Low Grades by Supervisor Gender (2006-2008)

To further illustrate the impact of the performance grades on employees, we look at the marginal return in bonus payments. Table 3.1 shows the

absolute differences in average bonus payments for different levels of the hierachy when grade 3 is assigned instead of grade 2 and grade 4 instead of grade 3. Level 1 is the lowest managerial level in this firm. It can be seen that for the lowest levels, receiving grade 2 instead of grade 3 already accounts for more than 3,000 Euro which equals about 4,5% of the total salary. This monetary difference is substantial. The table also shows that the marginal return for receiving grade 3 increases with the hierarchical level, in absolute and relative terms. Moreover, the marginal loss for receiving grade 2 compared to grade 3 is larger than the marginal loss when receiving grade 3 instead of grade 4 on every level. Assigning grade 2 to employees hence implies non-negligible monetary consequences making it even harder for supervisors to take this decision.

As a direct measure for differentiation in grades, the standard deviation of grades is computed per unit. The employee-level data is therefore aggregated on unit-level by means of the supervisor ID. Units with only one employee are excluded.[12] The result is shown in figure 3.4 which shows a higher standard deviation in grades for teams with a female supervisor compared to teams with a male supervisor.

The differences in the standard deviation are significant for the years 2007 and 2008[13]; this indicates that female supervisors differentiate more in their performance evaluations. However, as differences in other observables between male and female supervisors may drive the descriptive results, we will use regression analyses to further eliminate observed and unobserved heterogeneity of supervisors and work units.

[12] Additionally, we computed the coefficient of variation and the span of grades. In our view, the standard deviation is the most appropriate measure here also accounting for the frequencies of differentiated grades which the span of grades ignores. Moreover, for a scale from 1 to 5, a standardization by the coefficient of variation is not necessary. However, the figure is the same when using the other two differentiation measures.

[13] The test was significant on the 10% level in 2007 (p=0.0533) and on the 1% level in 2008 (p=0.0002).

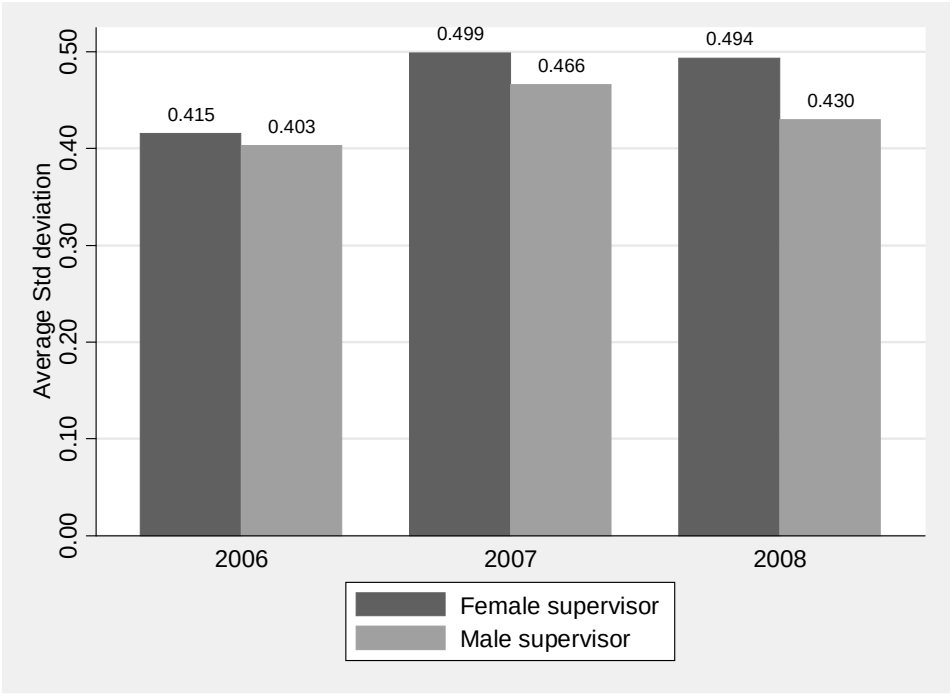

Figure 3.4: Standard Deviation of Grades within Workgroup by Supervisor Gender

3.5 Gender Differences in the Differentiation of Performance Evaluations

This section presents our empirical strategy and the regression results for gender effects in the differentiation of performance grades. In the first step, a work unit is defined as the group of employees who are evaluated by the same supervisor as described above (identified based on supervisor ID). The data is thus aggregated on supervisor level and the standard deviation of grades in the unit is the dependent variable.

The main explanatory variable is a dummy that takes the value 1 if the supervisor is a woman. A number of work unit and supervisor characteristics are computed as control variables because there are several reasons to believe that these may be correlated with the supervisor's gender and also have an impact on the differentiation in grades. It may well be that, for

example, the age of the supervisor positively correlates with a higher compression of ratings, because a higher conscientiousness due to career concerns can be assumed of younger supervisors. In addition, older and longer-tenured employees may have built up social ties with subordinates, making it more difficult for them to differentiate to the bottom of the scale. As the observed female supervisors are on average younger than their male colleagues, this may drive the result. Furthermore, the size of a workgroup can influence differentiation in grades, as in smaller teams, the personal contact between supervisor and subordinates is closer, rendering it more difficult for the supervisor to differentiate between the group members. Heterogeneity of the team, including the differences in status, might be an important influencing factor in the differentiation in grades, causing heterogeneity in grades to be driven by heterogeneity in status. Moreover, we want to control for further unit characteristics such as the share of newly hired, promoted and job movers which might influence the evaluations. Newly hired employees have not accumulated specific human capital yet which may results in lower performance evaluations compared to incumbents. In addition, newly hired supervisors may not have had a lot of time to observe the employees´ effort which can cause a higher compression in ratings. If newly hired supervisors are mainly male, this could drive the result.

Our regression model will account for these observable influencing factors. In particular, the work unit characteristics include the female share, average age of employees per unit, average hierarchical level[14], average salary (measured in Euro) and the average contractual working time of employees. In total, we observe nine levels for managerial employees where jobs on the lowest level (level 14) are the typical university-entry positions to young academics in this firm. The contractual working time has the maximum value 1 which stands for a full-time position. Furthermore, the percentages of newly hired, promoted and horizonally moved employees per work unit are computed. A further control variable on work unit level is the standard de-

[14]In this firm, the salary level is defined as a contractual salary band with lower and upper salary limits and determines the hierarchical level of an employee. In total, we observe 9 salary levels for managerial workers in the data set.

viation of employees‘ levels which is a work unit heterogeneity measure to control for objectively measurable variability of team members that may influence the differentiation in grades. Besides the gender of the supervisor, supervisor age, contractual working hours and the level of the supervisor are included. We also control for three additional variables that indicate whether the supervisor is newly hired, was currently promoted or changed business units. Descriptive statistics of all variables are presented in table 3.2.

	Mean	Std.dev.	Min	Max
Dependent variables				
Std. dev. of grades per unit	0.390	0.305	0	1.527
Fraction of grade 2	0.034	0.131	0	1
Supervisor variables				
Female supervisor (0/1)	0.098	0.298	0	1
Level supervisor	18.212	1.893	14	23
Working hours supervisor	39.889	1.150	20	47.4
Age of supervisor	47.065	6.511	24	64
Promotion of supervisors (0/1)	0.159	0.366	0	1
Newly hired supervisor (0/1)	0.011	0.103	0	1
Subdivision move of supervisor (0/1)	0.014	0.118	0	1
Work unit variables				
Female share	0.190	0.281	0	1
Contractual working time	0.983	0.057	0.4	1
Average hierarchical level	15.872	1.329	14	20.75
Average age	44.676	5.624	25	63
Number of unit members	3.867	2.832	1	21
Average salary (EUR)	83,389	17,118	30,092	157,800
Promotion rate	0.139	0.238	0	1
Rate of newly hired	0.030	0.115	0	1
Rate of subdivision moves	0.020	0.125	0	1
Std. deviation of salary levels	0.929	0.539	0	3.266

This table includes the descriptive statistics for the aggregated and pooled data (2006-2008) set on the supervisor level. Please note that standard deviations are only computed for units with at least two employees.

Table 3.2: Means and Standard Deviations

In the first step, we use a data set where the supervisor is the unit of observation. Therefore, we estimate our model based on pooled OLS and

random effects regressions. The regressions equation is as follows:

$$Std.dev.ratings_{jt} = \alpha + \beta \cdot FemaleSupervisor_{jt} + \gamma \cdot X_{jt} + \lambda \cdot Z_{jt} + \delta \cdot Y_t + \nu_{jt}$$

where j denotes the work unit in year t, X_{jt} covers general *workgroup characteristics,* which besides the above described variables includes business unit dummies. Z_{jt} contains the above discussed *supervisor characteristics.* Additionally, we control for year fixed effects with Y_t. Table 3.5 reports the results of the regression analysis. Column (1) and (2) show the OLS and the random effects OLS regressions respectively.

We again find that groups led by women have a higher standard deviation in grades compared to groups led by a male supervisor. Moreover, the age of the supervisor is negatively correlated with the standard deviation in grades, suggesting that younger supervisors differentiate more. The size of the team has a significant and positive coefficient. In addition, the standard deviation of salary grades in the workgroup show the expected positive significant relationship with the standard deviation in grades.

The analysis indeed suggests that female supervisors differentiate more between their employees. However, there might still be unobserved unit characteristics which may bias the results. To rule this out, we apply a second empirical approach that allows us to identify an effect based on within work unit differences with regard to the gender of the supervisor. For this approach, a workgroup is defined based on the employees who worked together over the whole period. In 2006, we assign to every employee who is working under the same supervisor a team ID. Thus, 2006 constitutes the base year of identifying a workgroup. The 2006-team ID of an employee is transferred to the remaining year-observations of the same employee, so that it stays constant for every employee over the whole period. We then identify a workgroup based upon the majority of employees of the original assigned team in 2006 who are still working under a common supervisor, regardless of whether it is the same supervisor as in 2006. If employees are not working under the same supervisors as their workgroup colleagues, they may have changed units or left the firm, so they are eliminated from the sample of the respective year.

	Pooled OLS	Pooled RE
Dependent variable:	Std. dev. of grades	
	(1)	(2)
Female Supervisor (0/1)	0.046**	0.049*
	(0.023)	(0.027)
Age of supervisor	-0.004***	-0.005***
	(0.001)	(0.001)
Working hours supervisor	0.002	0.005
	(0.007)	(0.006)
Size of team	0.016***	0.017***
	(0.002)	(0.002)
Share of promotions	0.023	0.003
	(0.034)	(0.032)
Share of new hires	-0.051	-0.045
	(0.076)	(0.073)
Share of horizontal moves	0.111	0.103
	(0.090)	(0.081)
Std. dev. of salary grades	0.034***	0.023***
	(0.013)	(0.014)
Controlled for promoted/ new/ moved supervisors	yes	yes
Observations	2,460	2,460
R^2	0.089	
Number of teams		1,144

Further control variables on supervisor level are salary grade and working time of the supervisor, team controls are share of women, average age, average salary level, salary and contractual status (full-time or part-time). Year business unit fixed effects are included. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 3.3: OLS and Random Effects Regression with Standard Deviation of Grades in Workgroup

When aggregating the data set on this workgroup level, the supervisor gender varies over time because supervisors move between workgroups. Now we can identify a gender effect by estimating workgroup-fixed effects regressions based on *within* workgroup difference in supervisor gender. The dependent variable is again the standard deviation in grades within the workgroup. The coefficient of the female dummy will reveal whether a female supervisor differentiates more in evaluations of the same workgroup compared to a male supervisor. As there is hardly any reason why the change of a supervisor to one of the opposite gender should be related to any unobserved time-variant workgroup characteristic, this approach should allow us to get very close to an exogenous variation in supervisor gender.

	Fixed Effects Regressions (OLS)		
Dependent variable:	Std.dev. of grades		
	(1)	(2)	(3)
Female Supervisor (0/1)	0.122**	0.119**	0.085
	(0.059)	(0.060)	(0.071)
Female Supervisor × Female Share			0.204
			(0.240)
Age of supervisor^2	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)
Working hours supervisor	0.005	0.003	0.005
	(0.018)	(0.019)	(0.189)
Team size	0.030***	0.030***	0.030
	(0.007)	(0.007)	(0.007)
Level dummies supervisor	yes	yes	yes
Other team characteristics[a)]		yes	yes
Observations	2,208	2,208	2,208
R^2	0.044	0.048	0.049
Number of teams	835	835	835

[a)] These include average age, female share, average level and salary, std. deviation in levels and average working hours. Year and business unit fixed effects are included. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 3.4: Standard Deviation of Grades - Workgroup Fixed Effects

For the analysis, control variables are the same as above, obviously adapted

according to the team composition of this approach. Due to the new data aggregation approach we cannot include the share of promoted, newly hired or horizontally moved employees as control variables because these employees are eliminated from the sample. Table 3.4 reports the results of the fixed effects regressions. In column (1) only supervisor controls are included whereas the specification in column (2) controls additionally for other team characteristics such as for example the standard deviation in levels and the female share. The coefficient of the female supervisor dummy is again positive and significant on the 5 percent-level in both specifications, which indicates a significantly higher standard deviation in grades for female supervisors. However, it might be that the degree of differentiation of a unit is influenced by the share of female or male employees. As former studies report a positive evaluation bias for same-sex dyads in subordinate-supervisor-relationships (Furnham and Stringfield (2001), Varma and Stroh (2001)), a female supervisor may evaluate differently when the majority of subordinates are male. To rule this out, we included the interaction term of the gender of the supervisor and the share of female employees in the fixed effects regression in column (3). The coefficient of the interaction term is not significant, so that there is no difference in differentiation only because of the gender composition. Overall, the variation from a male supervisor to a female supervisor of a workgroup with the same employees increases the standard deviation in grades. In a next step we will investigate in more detail the assignment of poor grades to subordinates.

3.6 Do Female Supervisors Give Poor Grades More Frequently?

Supervisors can differentiate toward the top or the bottom of the evaluation scale. As Jack Welsh (2003) for instance claims, it is especially tough to differentiate towards the bottom of the scale and nominate low performers in organizations. In his book he says: "Dealing with the bottom 10 is tougher(...). No leader enjoys making the tough decisions. We constantly

face severe resistance from even the best people in our organization"(Welch (2003): 198). Therefore, we here further investigate whether female managers are more likely to give poor grades to subordinates. Our empirical approach is similar to the one above. We start with aggregating the data on the supervisor level (based on supervisor ID). The dependent variable of our analysis is the fraction of employees per work unit that received the grade below the middle grade, namely "Partially meets expectations" (2). We chose this variable because supervisors might tend to put a poor performer in the middle category "Fully meets expectations" (3) hiding him or her in the mass of average performers. But supervisors who dare to give the 2 to subordinates demonstrate real competence in differentiating. OLS and random effects regressions are estimated. Work units with only one employee are again excluded. Control variables are work unit (X_{jt}) and supervisor characteristics (Z_{jt}) from above, so that we estimate the following specification:

$$PercentageGrade2_{jt} = \alpha + \beta \cdot FemaleSupervisor_{jt} + \gamma \cdot X_{jt} + \lambda \cdot Z_{jt} + \delta \cdot Y_t + \nu_{jt}$$

Table 3.5 reports the regression results. Both models show a significant and positive coefficient of the female supervisor dummy on the fraction of grade 2 in a team. A unit with a female supervisor has a 2 percentage points higher fraction of grade 2 evaluations than a unit with a male supervisor in the random effects regression. The coefficient is substantial, given that only about 3 percent of employees receive a poor grade in this firm. In a larger team, there is a slightly higher percentage of poor grades, which is consistent with the above finding of higher standard deviations in grades for larger teams. The percentage of promoted employees is negatively correlated with the fraction of poor grades, which suggests that promoted employees are often high performers who have a low probability of receiving a poor grade.

In addition, we use the data with the newly defined workgroups from above to eliminate unobserved heterogeneity between units.The same regression is conducted with the new dependent variable. Results are reported in table 3.6. Again, the coefficient of the female supervisor dummy is positive

	Pooled OLS	Pooled RE
Dependent variable:	Fraction of Grade 2	
	(1)	(2)
Female Supervisor (0/1)	0.017**	0.020**
	(0.007)	(0.010)
Age of supervisor	0.000	-0.000
	(0.000)	(0.000)
Working hours supervisor	0.000	0.001
	(0.001)	(0.001)
Size of team	0.001*	0.002**
	(0.001)	(0.001)
Share of promotions	-0.024***	-0.019**
	(0.009)	(0.009)
Share of new hires	-0.005	-0.011
	(0.019)	(0.016)
Share of horizontal moves	0.003	0.007
	(0.029)	(0.027)
Controlled for promoted/ new/ moved supervisors	yes	yes
Observations	2,460	2,460
R^2	0.044	0.041
Number of teams		1,144

Marginal effects reported in (1). Further control variables on the supervisor level are salary grade and working hours. Further team controls are average age, average hierachical levels salary and average contractual status. Year and business unit fixed effects are included. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 3.5: OLS and Random Effects Regressions with the Fraction of Giving Poor Grades

	Fixed Effects Regression (OLS)		
Dependent variable:	Fraction of grade 2		
	(1)	(2)	(3)
Female Supervisor (0/1)	0.033**	0.034**	0.043**
	(0.016)	(0.016)	(0.020)
Female Supervisor × Female share			-0.047
			(0.057)
Age of supervisor^2	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Working hours supervisor	0.006	0.007	0.007
	(0.005)	(0.005)	(0.005)
Team size	0.002	0.002	0.002
	(0.002)	(0.002)	(0.002)
Level Dummies Supervisor	yes	yes	yes
Other team characteristics[a)]		yes	yes
Observations	2,230	2,230	2,230
R^2	0.016	0.019	0.020
Number of teams	843	843	843

[a)] These include average age, average level, female share, average salary and average working hours. Year and business unit fixed effects are included. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 3.6: Fraction of Poor Grades - Workgroup Fixed Effects

and significant on the 5 percent level in both specifications. On average, in a work unit in which a male supervisor has been exchanged by a female one the fraction of evaluations with grade 2 increases significantly by about 3 percentage points. Given that the share of grade 2 is on average 2.9% in the observed teams, the effect is immense: a 100% increase in the share of grade 2 when the team is supervised by a women. Hence, female supervisors indeed show the evaluation behavior many firms strive to achieve: they dare to differentiate more and are less reluctant to give poor grades. In column (3) we again include the interaction of the female unit members and the gender of the supervisor. A higher share of low grades could be driven by the mere fact that the majority of subordinates are men, and female supervisors give lower evaluations to men compared to women. However, we see that the interaction term is insignificant which rules out this alternative explanation. Female supervisors indeed seem to be less reluctant to assign poor grades and are hence performing well in the tough part of evaluations.

3.7 Are Female Supervisors Indeed Good in Leading People?

So far, we found strong evidence that women evaluate in a more differentiated manner than men and that they give lower grades more often. In this section, we want to validate whether women in this firm are perceived by their managers as being better in leading their subordinates than men. In the observed firm, managers have to evaluate the leadership competencies of supervisors based on 7 leadership principles once a year. For every leadership principle, the supervisor receives a grade of 1 to 3 from his or her direct manager; 1 stands for "Development Need", 2 for "On Target" and 3 for "Strength". As revealed from company representatives, these evaluations are relevant for promotions and the career development of managers. One of the defined leadership principles is denoted as "Leading People" and depicts the ability of the supervisor to manage his employees in such a way that they contribute to the firm‘s value. The firm gives precise descriptions about the

competencies that are required for being strong in "leading people". Two of these competencies are "Giving open and honest feedback and fairly rewarding and acknowledging performance" and "Inspiring in others a passion to succeed, enabling and empowering them to achieve optimum performance." In our view, these descriptions are closely related to the objectives of differentiated performance evaluations, because more differentiation requires honesty and implies fair rewards for employees and enhanced incentives to increase effort. Also, we find a slightly positive and significant correlation between the differentiation in grades and the evaluation of "Leading People" (Spearmen 0.05, p=0.0613).[15] Thus, we here look at the differences in evaluations for the principle "Leading People" between male and female supervisors. There are many empirical studies in the management literature comparing overall performance grades of women and men (see Roth et al. (2011) for an overview) that show that women are on average better evaluated than men. However, we here investigate the evaluations of leadership principles of supervisors which is different to the overall performance grade. Not every supervisor in the firm receives the leadership evaluations yet, so we cannot observe the evaluations for each of the observed supervisors; this is because evaluation of leadership performance is still in the process of being launched for every supervisor in the firm. Overall, 62% of managers in the data set received the rating 2 for being "on target", 34% received the grade 3. For only 4% of the managers, there is development needed in the dimension of leading people. Figure 3.5 shows that in every observed year, a higher fraction of female supervisors is "Strong" in "Leading People" (receive the best grade 3) than male supervisors. Over all years, 45.6 percent of female supervisors receive the grade 3 compared to only 32.5 percent of male supervisors.[16]

We further conducted a simple, pooled regression to analyze whether women are more likely of being evaluated with grade 3 in "Leading People" when other supervisor characteristics are controlled for. The dependent variable is a dummy that takes the value of 1 for being strong in "Leading

[15] The overall leadership evaluation is also positively correlated with the standard deviation of grades per work group (0.07, p=0.001).

[16] Please note that we observe 291 (33) evaluations in 2006, 481 (73) in 2007 and 505 (85) evaluations in 2008 for male (female) supervisors.

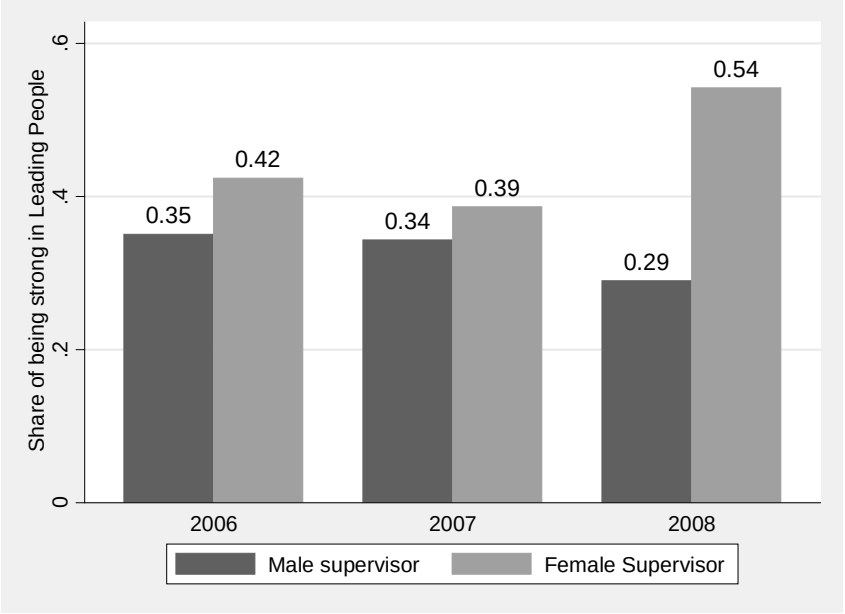

Figure 3.5: Fraction of Supervisors Receiving the Grade 3 ("Strength") in the Dimension Leading People

People". We use the aggregated data set on the supervisor level from above and control for age, salary level, working hours of the supervisor and size of the supervisor‘s team. Also, dummy variables for supervisors that are just being promoted, newly hired or changed jobs are added. In addition, we control for business unit and year fixed effects. Probit and random effects probit regressions are computed. Table 3.7 reports the results.

The first column reports marginal effects of the probit regression and shows that female supervisors are by 13.1 percent more likely of being evaluated with the best grade in the leadership dimension "Leading People", which is significant at the 5%-level. The random effects probit regression confirms the result. Overall, the results reinforces that female supervisors are strong in leading people which is in line with a higher differentiation between employees.

However, we also looked at the other leadership principles to reveal whether women also rate higher in other dimensions. Figure 3.6 reports the average

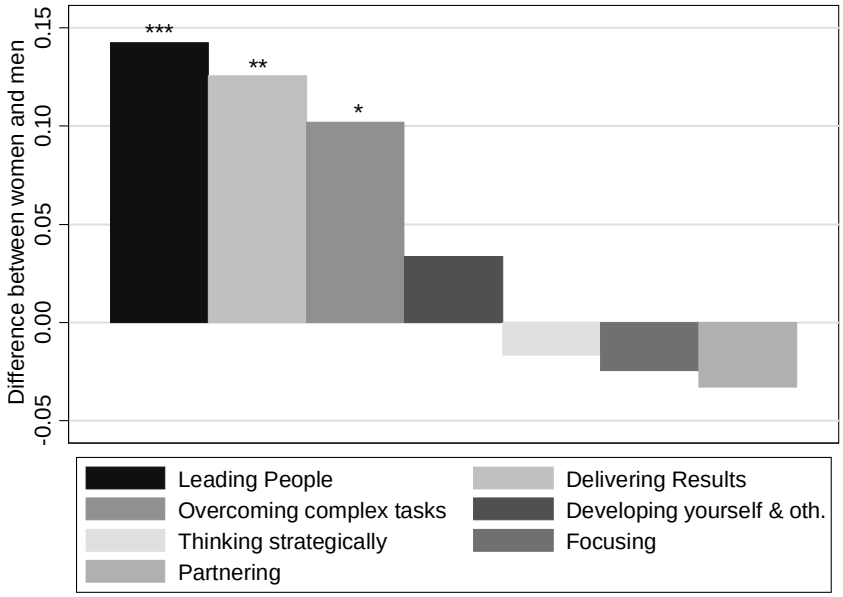

Figure 3.6: Overview of leadership principles - Difference of female to male supervisors

difference between female and male manager's leadership evaluations.[17] As we can see, the difference is the highest for "Leading People". Female manager's rate also significantly higher (see ** and * in the figure) in two other principles: Delivering Results and Overcoming complex tasks.[18] But it is obvious that female managers do not rate better in every principle on average. The difference to men is the most negative for "Partnering" which describes the ability to "cultivate an active professional network inside and outside of the organization". While this could be a result of women's exclusion from "old boy's networks" (Davies-Netzley (1998)) and their resulting diminished opportunities to network, there are also theories about gender differences in networking behavior (Friebel and Seabright (2011)). While women seem to focus on networks with "strong ties", men are rather focused on networks with "weak ties" which may be more effective at transmitting information according to Granovetter (1973). According to our data, women score indeed lower in "Partnering", but the difference is not significant. Overall, the comparison illustrates that the female managers seem to be especially strong in the dimension of leading other people.

3.8 Conclusion

In this paper, we addressed the question of whether female and male supervisors differ in their leadership competence to differentiate between their subordinates in performance evaluations. As the evaluations are tied to bonus payments, differentiation enhances the marginal return of effort for employees and implies an enhanced incentive setting (Bol (2011), Berger et al. (2010), Kampkoetter and Sliwka (2010)).

Our analysis was based on a unique firm-level data set of a large German

[17] Please note that *** $p<0.01$, ** $p<0.05$ and * $p<0.1$.

[18] "Delivering Results" is described with "Setting challenging yet realistic targets", "Pursuing goals with energy, drive and determination" and "Resolving conflicts, setting the right priorities, and allocating resources accordingly". "Overcoming complex tasks" is described as "Understanding how things get done in our complex environment", "Identifying how things could be done more simply and effectively and taking the necessary action" and "Challenging and eliminating activities that do not create value".

company for the years 2006-2008. With the information on the hierarchical connections in this firm, we were able to analyze the differentiation in grades together with supervisor characteristics of a work group. Based on pooled OLS regressions and panel regressions, eliminating unobserved heterogeneity of the work units, we observe women to differentiate more in performance evaluations than their male counterparts. Especially, female supervisors seem to assign poor grades twice as often as men. Based on survey data of these managers, we find evidence that this evaluation behavior is especially driven by women with children having a full-time working partner.

Differences in personality and social preferences between men and women may give explanations for the observed gender differences in evaluations behavior.[19] In detail, differences in reciprocity, in personality, or in lying aversion between men and women may possibly explain this observation. Women have empirically been shown to be more conscientious (Schmitt et al. (2008)) and more averse to lying than men (Dreber and Johannesson (2008), Conrads et al. (2011)) which may lead to more honest and therefore more differentiated performance evaluations. Also, a higher reciprocity of women can explain the results, as they have been shown to reward fair and punish unfair behavior more often then men (Eckel and Grossman (1996)).

Another explanation for our findings could be that we observe a (self-) selection of female managers that have proven to be good leaders. A selection of very able women in our data would imply that women may not be represented in leadership positions to the same extent than men. It might be that the firm would not equally value female and male competencies for leadership positions so that women have to show better qualities in order to be promoted to managerial positions. To check the relevance of higher differentiation for being a good leader, we see in figure 3.8 in the appendix that the standard deviation of grades assigned to subordinates increases with the hierarchical level in this firm.[20] This illustrates that differentiation is a

[19] Please see Croson and Gneezy (2009) for an overview of gender differences in risk attitudes and social preferences.

[20] Please note that the level 1 in the graph is one of the lowest levels of supervisors in this firm and equals level 16 from the abobe described levels. Only very few supervisors are assigned to level 14 or 15 which is why they are ignored here. Starting from level 4

relevant competence for being a leader in this company.

To rule out selection, we would need a measure for ability of women. But we are not able to measure ability of managers other than looking at the leadership evaluations. Although figure 3.6 suggests that women do not systematically score better in *every* leadership dimension, significant differences only occur for the principles in which women positively deviate from men. On average, female manager's hence seem to show better leadership qualities than men which could hint to (self-)selection of competent women. Although women rate higher in the observed leadership ratings, male managers may compensate the difference in other dimensions. For example, due to family responsibilities women may have on average shorter working hours than men. Based on a survey in this firm in 2010, it was revealed that female full-time managers work on average 1.2 hours less per week than male managers.

The observation that female managers show good leadership quality is consistent with management studies investigating gender differences with regard to leadership styles. For example, Eagly et al. (2003) show in their meta-analysis that women rate higher in the "contingent reward" dimension of "transactional leadership" than men. In accordance with the description of the "contingent reward" dimension, leaders with high scores in this dimension provide rewards for satisfactory performance, which has shown to improve employee's performance.[21] Here, women also score higher in the dimension "Leading People," which covers the competence of giving honest feedback and rewarding fairly.

A major limitation of our study is that we only use data of one single firm which may limit generalizability of the result. The same analysis should be conducted with similar data sets from other companies to strengthen the result. In addition, this study is not able to give explanations for the observed differences, so we cannot definetely consider gender differences in personal preferences as the driving force of the result. Further research might be

the standard deviation in grades is significant to the former levels on the 1% level (also in regressions when controlling for influencing variables).

[21] An exemplary item for the contingent reward dimension from the Multifactor Leadership Questionnaire (MLQ) asked to subordinates is: "My supervisor works out agreements with me on what I will receive if I do what needs to be done."

encouraged to obtain a deeper understanding of the different evaluation behavior of men and women. We aimed to show empirically, based on a rich personnel data set, that women perform well in one important dimension of being a leader: in setting incentives for employees and efficiently leading people. Our data even shows that they outperform male supervisors in this leadership dimension which is often considered as a key competency of managers.

3.9 Appendix of Chapter 3

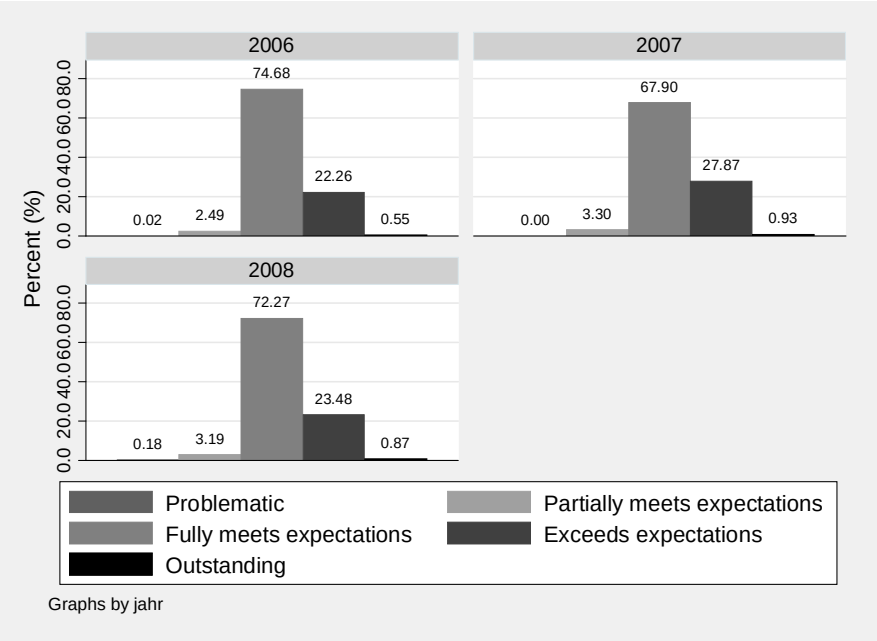

Figure 3.7: Distribution of Performance Evaluations by Year

	Pooled Probit	Pooled RE - Probit
Dependant variable:	Strong in leading people	
	(1)	(2)
Female Supervisor (0/1)	0.131**	0.582**
	(0.053)	(0.267)
Age of supervisor	-0.014***	-0.062***
	(0.003)	(0.016)
Size of team	0.018***	0.084***
	(0.005)	(0.027)
Observations	1,055	1,055
R^2	-636.697	-594.754
Number of teams		628

Marginal effects reported in column (1), not (2). Additional control variables are age, salary level, contractual working hours of the supervisor and the size of the team. Robust standard errors in parentheses. Year and business unit fixed effects are included. *** p<0.01, ** p<0.05, * p<0.1

Table 3.7: OLS and Random Effects Regression of Being Strong in "Leading People"

	2006	2007	2008	Total
Female Supervisor (Mean/ Sd)	3.17 (0.50)	3.30 (0.56)	3.22 (0.56)	3.23 (0.54)
Male Supervisor (Mean/ Sd)	3.21 (0.47)	3.26 (0.52)	3.22 (0.51)	3.23 (0.50)
Number of observations	4016	4177	4923	13116
Evaluations by female supervisor	8.2%	8.9%	9.5%	8.9%

Table 3.8: Mean and Standard Deviations of Grades by Supervisor Gender

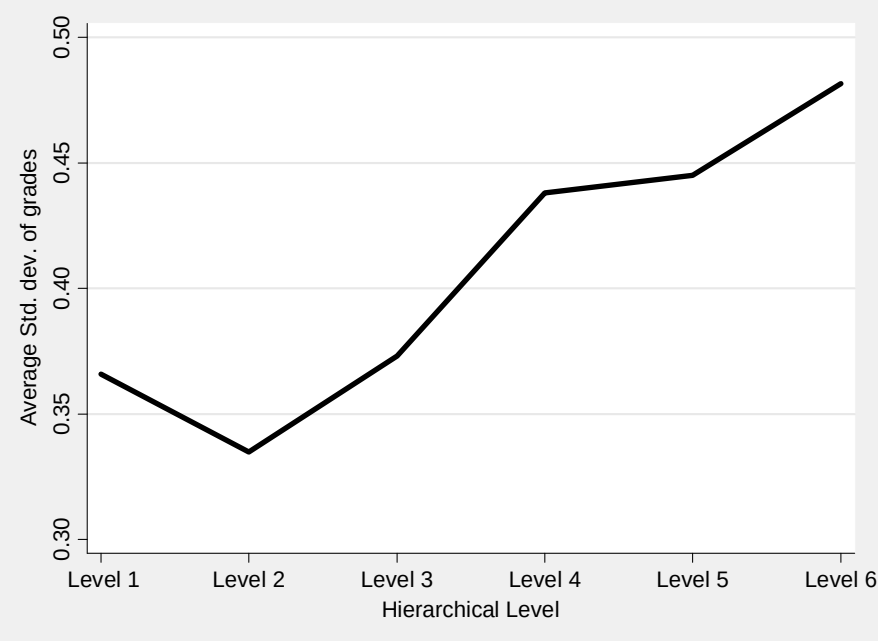

Figure 3.8: Average Grade Differentiation per Unit by Hierarchical Level of Supervisor

Chapter 4

Determinants and Effects of Target Agreement Systems: An Empirical Investigation of German Firms[1]

4.1 Introduction

Goals can regulate motivation and increase performance (Locke (1968)). So far, an extensive body of psychological studies has investigated the effects of goal-setting with robust insights: Goal setting may increase performance because it directs the attention and action of individuals and encourages overall effort (Latham and Locke (1990), Locke and Latham (2002)). Today, many organizations incorporate goal setting in incentive schemes and other human resource activities. As Hale and Whitlam (1998) say: "Target setting might potentially impact upon a very large number of organizational or human resource initiatives and policies" (p.50). Besides directing employee actions and stipulating standards for performance, goals can be linked to remuneration systems such as bonus payments. Empirical evidence suggests that goals increase performance and efforts of individuals. But can target

[1]This chapter is based upon Breuer and Zimmermann (2011).

setting for employees improve organizational performance? With this paper, we empirically address this question by analyzing a large-scale survey data set of German firms.

So far, the effects of goal setting has intensively been studied in the context of *individual* cognitive processes and behavior. Locke's "goal-setting theory" has shaped a body of psychological research in which goals are seen as a "reference standard for satisfaction versus dissatisfaction" (Locke and Latham (2002): p. 710), so that exceeding the goal provides increasing satisfaction while falling short of the goal raises dissatisfaction with increasing discrepancy. Many studies in cognitive psychology revealed that, compared to easy and/or vague goals that urge people to "do their best", goals that are specific and challenging are more likely to boost individual performance (see Locke and Latham (2002) for a literature review). Following the logic of goals as references points, Heath et al. (1999) modeled goals as the subject of the value function in Kahneman and Tversky's (1979) prospect theory with loss aversion regarding goal achievement. Also, goals serve as performance standards in non-linear incentive contracts such that when a standard is attained, a "target bonus" is paid out (Murphy (2001), Rablen (2010)). Some economic studies analyzed how individuals may set goals for themselves addressing self-control problems or the selection of reference standards (Falk and Knell (2003), Koch and Nafziger (2011), Koch and Nafziger (2009), Kaur et al. (2010)). In addition, some recent empirical studies have shown further interest in goals with the focus on how employees' targets are set and influenced by supervisors or firms (Bol et al. (2010), Anderson et al. (2010)).

Despite the work about the motivational effects of goals on individuals, rarely the impact or use of target setting for employees has been addressed on the organizational level. But as implementing systematic target setting for employees in organizations is costly, the benefits have to be evaluated. Benefits may also vary substantially with the character of a firm. In this paper, we first aim to investigate attributes of firms which are most likely to benefit from target setting for employees. Second, by making use of panel data methods we analyze whether the introduction of target agreements for employees show an effect on organizational productivity.

Somewhat related to our paper are the early management studies on the concept *management by objectives* (MBO) (Drucker (1954)) which look at the productivity gain of MBO on organizational level. The evidence suggests that MBO positively affects organizational performance (see Rodgers and Hunter (1991) for a meta-study). However, MBO includes goal setting as only one of three components. These are: participative management, goal setting and objective feedback (Rodgers and Hunter (1991)). Additionally, the concept of MBO is more holistic in the sense that objectives are set on firm level and translated into individual targets down the hierarchy. Mostly MBO was studied based on case studies and small cross-sectional samples limiting the generalizibility of results (Rodgers and Hunter (1991)). In our analysis, we focus on the usage of *target setting* for employees and investigate the organizational determinants and effects with large-scale panel data.

Our analysis is based on a unique data set from the German Institute for Employment Research (IAB) Survey, which covers a representative sample of more than $16,000$ German firms. The survey includes questions about employment characteristics and management practices, focusing on different topics over the years. In 2005 and 2007, the survey asked company representatives whether their firm uses target agreements for managing employees. In 2005 and in 2007, 24.1% and 26.9% of the firms reported to use the practice. We additionally observe many other firm characteristics such as size, workforce structure and change policies which are included in regression analyses.

The paper is organized as follows: Section 4.2 derives hypotheses about the firm characteristics that determine the use of target setting for employees and about the performance effect. After a description of our data set in section 4.3, our empirical strategy is described in 4.4. The hypotheses are empirically tested in section 4.5. Section 4.6 concludes.

4.2 Determinants of Target Setting and Performance Effects

In what follows, we derive several hypotheses about workforce and firm characteristics that may determine whether an organization uses target agreements for employees. As mentioned above, targets may serve as a performance standard for employees against which their performance is measured in the performance appraisal process (Hale and Whitlam (1998)). In our sample, about 75 percent of the firms that use target setting for employees also use performance evaluations. Hence, determinants of using target setting are related to those of using performance evaluations that have been studied by Brown and Heywood (2005) and Addison and Belfield (2008).[2] The degree of target achievement might be used as a performance measures for promotion or payment decisions. Arguments about which firms may be more likely to use the practice are hence partially derived by looking at who will most probabily benefit from incentive schemes. We have separated the relevant firm characteristics in workforce and structural firm characteristics.

Workforce characteristics

We expect that firms with relatively low employee tenure are more likely to use target setting for employees. Low-tenured employees may have a higher need to be steered toward the objectives of the firm than those with longer tenure. In addition, target agreements enable managers to define specific weaknesses that the employees have to improve for future job fulfillment, which is much more important for employees early in their career than for those later in their careers. Target achievement is also a kind of performance measures which facilitates looking at low-tenured employees‘ abilities to assign them to the right jobs (Jovanovic (1979)). Moreover, linking target attainment to bonus payments is especially important for lower-tenured employees, as employees with higher tenure are incentivized by deferred compensation, which raises the cost of dismissal for low effort (Lazear (1979), Lazear

[2]In addition, Grund and Sliwka (2009) have analyzed individual and job-based determinants of performance appraisal usage in Germany.

(1981)). In our data, we do not directly observe the average employee tenure. Therefore, we use the average turnover rate and the percentage of employees with temporary contracts as proxies for tenure such that the higher the employee turnover in a firm and the higher the percentage of employees with temporary contracts, the lower the average tenure of employees.

H1: Firms with a high average tenure are less likely to use a target agreements for employees than are firms with a low average tenure.

We expect an increased probability of using target agreements for firms with more qualified workers compared to firms with low qualified workers. Since the profit loss of low effort by highly qualified workers is higher than for low qualified employees, firms with a more qualified workforce should be more willing to set up target agreements. In addition, highly qualified employees are typically assigned to non-standardized white-collar jobs with a higher level of discretion over their tasks. A higher level of job discretion increases the need of directing employees‘ attention to the firm's objectives. We use three variables to measure the qualification of employees: the percentage of skilled workers (i.e., workers fully trained in their jobs), the percentage of workers with a university degree and the supply of written personnel development and training plans to workers in the firm, since a firm with a formal personnel development is likely to employ more qualified employees.

H2: Firms with a higher share of qualified employees are more likely to use a target agreements for employees than are firms with a lower share of qualified employees.

Structural firm characteristics

Several non-workforce related characteristics may be relevant for the application of a target agreement system. We expect the size of a firm, measured by the size of the workforce and by the firm‘s sales volume, to be a determinant of the use of target agreements. First, larger firms profit from economies of scale when they use costly human resource practices. Second, firms with higher monetary gains have more monetary resources with which to install systematic target agreements for employees, particularly when goal

attainment is tied to bonus payments. Moreover, since the transparency of individual effort contribution decreases with firm size, the likelihood of shirking is increased in larger firms. Setting targets for employees may help to reduce shirking opportunities.

H3: The size of a firm is positively related to the use of a target setting for employees.

We assume that firms that have recently experienced reorganizations are more likely to use target agreements with employees. Restructuring may lead to conflicts because of changes in team compositions, supervisors (Howard and Frink (1996)), and tasks. Target agreements, hence, can help to direct the attention of employees to their new job responsibilities and away from contextual problems of the restructuring.

H4: Firms that have recently undergone job reorganizations are more likely to use target setting for employees than are firms that have not.

Another important determinant may be the level of labor union organization. We expect collective agreements and the presence of a works council to be negatively correlated with the use of target agreements. Target attainment is frequently linked to payment schemes that may violate wage schemes pursued by unions. In Germany, works councils have co-determination rights in most personnel-related topics, leading to negotiation costs. To avoid these costs, labor-union-organized firms may choose not to apply target agreements for their employees.

H5: Collective agreements and the presence of a works council are negatively correlated with the probability of a firm's use of target agreements with employees.

Performance effect of introducing target agreements

If goals or targets increase individual productivity, an organization that implements targets for employees is expected to profit from target setting on a corporate level. In cognitive psychology, four mechanisms have been empirically identified through which goals affect individual performance (see

Locke and Latham (2002) for an overview). Goals direct the attention of individuals toward relevant activities and away from irrelevant activities; goals act as energizers, such that for both physical and cognitive tasks lead to greater effort than lower goals do; goals affect persistence, such that individuals continue exerting effort until they have attained the goal; and goals activate cognitive and task-relevant knowledge so the individual can manage the situation successfully (Locke and Latham (2002)). If targets are introduced as the standard of performance in incentive schemes (Murphy (2001), Hale and Whitlam (1998)), employees are induced to exert higher effort. Besides having a direct incentive effect on employees, the introduction of compensation-relevant targets may cause a sorting of employees (Cadsby et al. (2007)). While high-performing employees self-select into the organization because they can expect to achieve the target level of performance, low performers may be dismissed or leave the firm for one without target-dependent compensation arrangements. Overall, a performance increase for firms that implement target setting for employees can be expected.

H6: The introduction of target agreements for employees leads to improved corporate performance.

So far, empirical evidence that investigates the overall effect of the introduction of target setting for employees is limited. We only know about one further related study of Terpstra and Rozell (1994) who analyze the relationship of applying goal setting theory in firms and organizational profitability based on a survey data set. The focus of analysis is different as they ask firms' managers for the relevance of Locke's goal setting theory for their overall human resource practices while we focus on the application of target agreements to manage employees. Moreover, while they look at correlations in a cross-sectional design, we focus on the impact of the target setting implementation in a longitudinal research design.

4.3 Data Set

Access to the IAB Establishment Panel, Wave 2005 to 2008, was provided via remote processing by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB).[3] The IAB Establishment Panel is an annual representative survey of German establishments on topics like the determinants of labor demand. The survey has been conducted since 1993 in West Germany and since 1996 in East Germany. Since 2001 more than 16,000 organizations have been interviewed every year. Only those establishments that are mainly responsible for the operative business of a firm are included - for example, holding divisions are excluded - so a relevant data basis for employment-related topics can be assured. The main feature of the data set is the longitudinal format, which allows the use of panel data methods that control for unobserved heterogeneity among establishments. In 2005 and 2007, the establishments were asked if they conduct written target agreements with employees, a question that constitutes the main variable of our analysis. The exact question, translated into English, is, "Does your company use written target agreements for employees?". The regression analysis is primarily based on the pooled cross-sectional and longitudinal data from these two years. Table 4.2 shows that, in total, we observe 15,875 firms in 2005 and 16,004 firms in 2007. The companies in our sample employ on average 161 employees with a median value of 26 employees, illustrating that companies of all firm sizes are covered in the data set (see Table 4.1). In the next section we will describe the relevant variables for the analysis and the empirical strategy in detail.

4.4 Empirical Approach

The empirical strategy is as follows. To analyze the determinants of target setting we conduct binary probit estimations. The dependent variable y_{it} is a dummy that takes the value 1 for an establishment i if the company has

[3]Further information about the data set can found in the methodological report of the FDZ 01/2008 (Fischer et al. (2008)).

a target agreement system in the year t. Otherwise, the value is 0. As the question referring to the application of target agreements for employees was asked only in 2005 and 2007, the pooled probit estimations cover the two years while the standard errors are clustered for firms. Based on the hypotheses, we use several independent variables. To approximate the average *tenure* of a company, we use the turnover rate and the percentage of temporary workers of a firm. Based on these variables we compute two dummies, each of which takes the value 1 if the rates of an establishment lie below the 25th-percentile or above the 75th-percentile of the distribution in the relevant industry and year. These variables take into account the industry benchmark in order to facilitate interpretation of the effects.[4]

In 2007, the survey included ten items that asked for organizational changes in the preceding two years of which we use three items to measure *job reorganizations* in a firm. These three items indicate a change in work organization and asked for the "reorganization of departments or functions," the "relocation of responsibilities to the bottom," and the "introduction of group work or autonomous workgroups". Respondents answered with yes if there was the respective organizational changein the firm and no if not. Hence, three dummy variables are built, each of which took the value 1 if the establishment was affected by the change and 0 otherwise.

To approximate the *qualification level of employees*, we use three variables: the percentage of qualified employees, the percentage of employees with university degrees, and the use of formal personnel development or training plans. A professional personnel development is especially applied in firms with a substantial amount of white-collar employees which tend to be better educated than blue-collar workers. In the regression, the dummies for the percentage of qualified employees and those with university degree were used while these are again based on the respective variable distribution per industry and year. The existence of formal personnel development or training plans was measured based on a dummy variable taking the value 1 if respondents answered with yes and the value 0 for no.

[4]We also computed the regressions using the simple shares as dependant variables and do not find a qualitatively different result.

The *size* of the observed establishments was measured by the number of employees employed[5] (in 1000) and the inflation-adjusted revenue per year (in million EUR). We add the quadratic terms because we expect a decreasing marginal probability of using target agreements for increasing firm size. Of course, information on revenue was provided only for commercial organizations.

Unionization of a firm is measured by three dummy variables taking the value 1 for the existence of a works council, an collective agreement in the industry, or a collective agreement on the firm level and 0 otherwise. The descriptive statistics of all the dependent variables are shown in Table 4.1. All of the described workforce (x) and structural firm characteristics (z) may affect the probability of using target agreements (y) for employees so that:

$$P(y_{it} = 1|x_{it}, z_{it}) = G(\alpha + x'_{it}\beta + z'_{it}\delta + c'_{it}\gamma)$$

describes our empirical model. c'_{it} is a vector of further control variables and includes year, industry and legal form dummies. β and δ measure the effect of the independent variables on the probability of using target agreements for employees and α is the intercept. Because revenue was applicable only to commercial establishments, we conduct regressions with two separate samples: one sample that includes all establishments and one that excludes non-commercial establishments.

For analyzing the performance effect following an introduction of target setting, we use an OLS first-differencing approach to eliminate unobserved time-constant differences between firms. From the initial sample we only consider firms that had not installed target agreements for employees in 2005. Comparing the revenue growth of firms that introduced the system between the two years to the performance of those who did not, provides the opportunity to get closer to a causal relationship in the multivariate regression. Two performance measures are used as dependent variables (r_{it}): the log revenue for the observed year and a subjective evaluation of the revenue growth. In the survey, total revenue is reported for the previous year, so we take the rev-

[5] This number equals the personnel endowment at 31 July every year.

enue variables from the subsequent years, 2006 and 2008.[6] We consider for the analysis only those firms that previously stated that they use *revenue* as their transaction volume. Thus, banks, insurances, and non-profit-organizations are excluded, which view performance in terms of the balance sheet total, total premiums, or total budget. The *subjective evaluation of growth* is based on the following question in the survey: "Which development in business volume/ revenue do you expect for the current year (...) in comparison to the previous year (...)?" Answer possibilities were "Increasing", "Stay the same" and "Decreasing". The constructed dummy variable takes the value of 1 for an increasing revenue and 0 otherwise. Table 4.1 reports the descriptives for the revenue variables. The correlation of the two variables is quite high with 0.7 ($p<0.01$). A first-differencing approach requires to compute the differences of all variables between 2005 and 2007. The main explanatory variable hence takes the value 0 ($y_{i2007} - y_{i2005} = 0_{i2007} - 0_{i2005}$) if target agreements are not introduced or the value 1 ($y_{i2007} - y_{i2005} = 1_{i2007} - 0_{i2005}$) if they are introduced. Firms that use target agreement for employees in both years (so that $y_{i2007} - y_{i2005} = 1_{i2007} - 1_{i2005} = 0$) are ignored in the analysis. We further control for differences in the above workforce and firm-specific characteristics between the two years. The reorganization variables are not included as they are only observed in 2007. In the first-differencing OLS regression, all time-constant firm influence variables such as industry, legal form and federal state are eliminated (firm fixed effects) so that only the time-variant, unobserved heterogeneity remains. The estimated linear equation is:

$$\Delta r_i = \alpha + \beta \Delta y_i + \Delta \mathbf{c}_i' \boldsymbol{\gamma} + \Delta \upsilon_i$$

$\Delta \mathbf{c}_i'$ represents a vector of control variables and $\Delta \upsilon_i$ the first difference of the time-variant error term.

[6] The exact question from the survey was, "What was your transaction volume in the previous financial year?".

Variable	Obs.	Mean	Std.dev.
Key variable			
Target agreements (1/0)	31879	0.255	0.436
Tenure variables			
Turnover rate	31879	0.069	0.655
Share of temporary workers	31879	0.059	0.147
Qualification variables			
Share of qualified workers	31879	0.656	0.295
Share of workers with university degree	31879	0.099	0.192
Personnel development (1/0)	31879	0.282	0.450
Size variables			
Number of employees	31879	161	25.798
Revenue (in million)	17669	32.385	426.707
Reorganization variables			
Reorganization of departments (1/0)	16004	0.163	0.394
Relocation of responsibilities (1/0)	16004	0.112	0.338
Introduction of group work (1/0)	16004	0.061	0.258
Unionization variables			
Works council (1/0)	31879	0.325	0.469
Industry tariff commitment (1/0)	31879	0.421	0.494
Company tariff commitment (1/0)	31879	0.073	0.261
Other variables			
Expectation of revenue growth (1/0)	23887	0.366	0.431
Commercial firm (1/0)	31879	0.842	0.365
Industry Dummies (1/0)			
Agriculture & Mining	31879	0.041	
Manufacturing	31879	0.241	
Construction	31879	0.084	
Retail	31879	0.143	
Tourism & Transportation	31879	0.040	
Financial services & Insurances	31879	0.028	
Public & Private Services	31879	0.334	
Others	31879	0.086	

The table shows the pooled statistics for 2005 and 2007. The number of obs is reduced for the revenue variables because only commercial firms report this figure. The reorganization variables have only been collected in 2007 which reduced the number of obs.

Table 4.1: Descriptive Statistics

4.5 Results

In what follows, we will first present some descriptive statstics on the use of target agreements in German firms and report some descriptive evidence towards the hypothesized relationships. Subsequently, the regression analyses are presented.

4.5.1 Descriptive Results

The number of firms that use target agreement systems is shown in table 4.2. The percentage of firms with target agreements increased by about 2.8% from 24.1% in 2005 to 26.9% in 2007. When computing the share of firms that use target agreements *and* performance appraisal in 2007 (see table 4.2), we see that 74.6% (3,214 out of 4,311) of the firms with target agreements probably combine it with performance appraisals. This supports the often discussed usage of target achievements as performance measure in appraisals.

		2005	2007	Total
Use of target agreements	Frequency	3,820	4,311	
	Percentage	24.1%	26.9%	
Additionally use of performance appraisal	Frequency	2,694	3,214	
	Percentage	17.0%	20.1%	
Number of firms		15,875	16,004	31,879

Table 4.2: Frequency of the Use of Performance Appraisals and Target Agreements

For a first indication regarding the hypotheses that cover the determinants of target agreement systems, we look at the mean values of the explanatory variables in firms with and without target agreements. Table 4.3 shows the averages of the company characteristics based on whether they have a target agreement system or not. The average turnover rate of firms that have installed target agreements is lower than that for firms without the instrument, which is in contrast to the expected relationship because we expected firms a lower average tenure and hence a higher average turnover to be more likely to use targets for employees. But the average percentage

of employees with temporary contracts shows the expected pattern. In firms with target agreement systems, the percentage of employees with temporary contracts is higher (6.8%) than in firms without these (5.6%). As suggested in hypothesis 2, firms with target agreements have a higher average percentage of qualified (75.6%) and graduated employees (15.4%) than do firms without target agreements (62.1% and 8.0% respectively).

The number of employees indicates that companies that use target agreements are larger in size which is in line with hypothesis 3. While firms with target agreements employ on average 376 employees, it is only 88 employees for firms without it. Also, revenue is ten times as large for firms with the target setting compared the ones without.

The relevant variables for hypothesis 4 also show the expected relationship in the descriptive overview. The variable *job reorganization* takes the value of 1 if any of the three described job reorganizations were conducted in a firm. More than half (51.2%) of the companies with target agreements faced at least one of the three job reorganization in 2007, whereas only 18.1 percent of the companies without target agreements did. The same picture evolves when considering the single dummies for reorganizations of work.

Hypothesis 5 proposed that a negative relationship between collective agreements and the installation of appraisal systems can be expected, which is not obvious in the descriptive overview. Also more than a half (54.4%) of companies that use target agreements are bound to industry tariff commitments, while only 38 percent of the companies without target agreements are tied to industry tariff commitments. In summary, there is descriptive evidence for hypotheses 1-4, but the descriptive results are opposite to what was expected in hypothesis 5. However, this result may be driven by firm size as larger firms are more likely of being unionized. In a multivariate analysis, other observable and influencing factors will be hold constant.

Hypothesis	Firm characteristics	Without target agreements	With target agreements
H1	Turnover rate	7.5%	5.2%
	Percentage of temporary employees	5.6%	6.8%
H2	Percentage of qualified employees	62.1%	75.6%
	Percentage of employees with university degree	8.0%	15.4%
	Personnel development (0/1)	14.9%	67.3%
H3	Number of employees	88	376
	Net sales	10,410T€	104,584T€
H4	Job reorganization (0/1)	18.1%	51.2%
	Reorganization of departments (0/1)	11.7%	39.3%
	Relocation of responsibilities (0/1)	9.1%	24.0%
	Introduction group work (0/1)	4.3%	14.9%
H5	Works Council (0/1)	22.8%	61.1%
	Collective Agreement - Industry (0/1)	37.9%	54.4%
	Collective Agreement - Firm (0/1)	5.6%	12.5%

The calculations are based on pooled data from 2005 and 2007 (N=31,879) for all variables except the reorganization variables (N=16,004). The means are calculated for firms with and without target agreements in use.

Table 4.3: Mean Values of Company Characteristics for Firms With and Without Target Agreements

Performance measures	All	No introduction of TA	Introduction of TA
Average net sales	10.69 mill.€	8.00 mill.€	28.83 mill.€
% Sales growth	8.68%	7.28%	11.31%
Sales per FTE	0.12 mill.€	0.11 mill.€	0.17 mill.€
Positive profit expectation	5.5%	5.3%	6.4%

Only firms that do not use target agreements in 2005 are considered. Sales are taken from the 2008 survey as firms answer retrospectively.
Abbreviations: FTE - Full-time equivalent, TA - Target Agreements

Table 4.4: Performance Measures in Dependence of Having Introduced Target Agreements

In Table 4.4 we look at the descriptives of hypothesis 6. Out of the firms without target agreements in 2005 for which we also observe the information on target agreements 2007 (n=11, 413), about 11.4% (n=1296) of firms have implemented target agreements by 2007. Table 4.4 presents the means of several performance measures of 2007 based on whether the firm introduced target agreements between the years or not. On average, firms that implemented target agreements show significantly higher net sales in 2007 than firms that did not. While the sale growth ($\frac{sales_{2007}}{sales_{2005}} \cdot 100$) is on average 7.3% for firms without the introduction of target agreements, it is about 4% higher for firms that introduced target agreements. Moreover, the sales per full-time equivalent (FTE) is about 1.5 times larger for a firm that introduced targets. In addition, a slightly higher percentage of representatives in firms that introduced the system expect higher sales growth than firms without it. Overall, the descriptive results are in line with the hypothesis.

4.5.2 Determinants of the Use of Target Agreements

Table 4.5.2 shows the marginal effects of the probit regression with a dummy for the use of target agreements as the dependent variable. In the first two columns, all firms and commercial firms in both years are considered, while in column (3) and (4), only regressions based on 2007 are presented because of the inclusion of the reorganization dummies, which were available only for 2007.

The coefficients of the dummy for the lowest turnover quartile is significantly negative. Establishments with very low turnover compared to industry averages are less likely to use target agreements. This result is in line with the prediction and robust over all specifications. However, the highest turnover quartile does show the expected sign, but no significance. In addition, having a low share of employees with temporary contracts (below the 25th% quartile) in the industry shows significant negative coefficient of about -3.7 % in (1) supporting the suggested relationship between tenure and use of target agreements. This result is robust in nearly all specifications. The 75th%-quartile dummy for the percentage of employees with temporary contracts in the industry is positive and significant only in the first specification; while this result is in line with the hypothesis, it suggests that the effect is driven by firms with very low turnover rates and low percentages of employees with temporary contracts. Firms with a high share of high-tenured employees seem to have less need to set targets for performance. Employees in these firms might be already matched to well-suited jobs, may have proven to be aligned with the firm's objective and may be motivated by deferred compensation schemes.

The analysis of hypothesis 2 confirms the suggested relationship. Firms in the lowest quartile of the share of qualified employees are significantly less likely to have implemented target agreement systems than are firms with a medium level of qualified employees, which is robust over all specifications. However, the coefficient of the 75th%- quartile dummy for the percentage of qualified employees does not show the expected effect, leading to the conclusion that firms with an extremely low percentage of qualified employees in the same industry refrain from using target agreements. For them, the benefit from target agreements is limited because of their relatively low costs for lower effort by employees. The quartiles for the percent of graduate employees show the expected relationship. Firms that employ a percentage of graduate employees that lies above the 75th%-percentile are even more likely to have implemented target agreements which is robust in all specifications.[7]

[7] To rule out multicollinearity of the share of graduated and qualified employees, we computed the correlation between the variables which is $r = 0.33$ and thus not critical.

Furthermore, the coefficient of the personnel development dummy shows a highly significant and positive sign. With a 32.9 percent higher probability of using target agreements for employees when a personnel development exists, the effect is substantial and confirms the hypothesis.

As for hypothesis 3, the first regression shows a significantly positive coefficient of the number of employees, but the effect is surprisingly small with 1.3 percent increased probability of using target agreements per $1,000$ employees. Net sales were also used to indicate size, but the coefficients were far from significance and close to zero. This result is surprising, given that other studies have found that performance appraisals, which are typically combined with target agreements tend to be used in large firms (see, e.g. Grund and Sliwka (2009), Brown and Heywood (2005)). However, other covariates may be positively correlated with the size of the firm and hence capture the relevant correlations with the use of target agreements for employees.

Coefficients for the reorganization measures (hypothesis 4) can be estimated only in the 2007 specifications. All coefficients show a highly significant and positive sign for both samples. Firms that experienced reorganization, while controlling for size of the firm and other characteristics, were more likely to use target agreements. The introduction of group work shows the largest positive relationship, with a 9.0 percent higher probability of using target agreements.

According to hypothesis 5, the presence of a works council and collective agreements should be negatively related to the use of target agreements. However, as has already been indicated in the descriptive statistics, the opposite is found here. The presence of a works council is significantly positively correlated with the probability of using target agreements. One explanation for the positive correlation is that a works council may serve as a communication tool between the firm's management and employees (Freeman and Lazear (1995)). Due to information assymetries employees may generally distrust what management says. But the legal requirement that management has to disclose information to elected works council representatives may increase employees' trust in the firm, so that the implementation of new practices, such as a target agreement system, can even be facilitated. In addition, the

use of target agreements may suggest the presence of a formal performance evaluation process, and formality itself may be in the interest of the works council. Company tariff commitment also has a positive and significant relationship with using target agreements for employees but is not as robust as the works council effect.

In summary, relationships with the use of target agreements for employees are confirmed for firms with long-tenured employees, highly qualified employees, and job reorganizations. There is no size effect, and unionized firms are even more likely to use target agreement systems, which is contrast to the hypothesis. The results are the same when using continuous variables for the turnover rate, the share of temporary contracts and the qualification variables instead of the quartile dummies. The corresponding regressions are reported in Table 4.8 in the appendix.

To rule out the argument that the coefficients of determinants are driven by a specific year observation, we conducted the same regressions by including the three-year moving averages of the independent variables as a robustness check. The averaged variables are supposed to be a smoothed and more robust indicator of the underlying workforce or structural firm characteristic. The results are robust for the adapted independent variables (see table 4.9 in the appendix).

		Pooled Probit Regressions			
	Dependent variable:	Target agreement (0/1)			
		(1)	(2)	(3)	(4)
		all	comm.	all - 2007	comm. - 2007
Tenure	Turnover < 25th% percentile	-0.039***	-0.050***	-0.028***	-0.035***
		(0.007)	(0.009)	(0.010)	(0.013)
	Turnover > 75th% percentile	0.006	-0.004	0.010	-0.003
		(0.007)	(0.009)	(0.010)	(0.013)
	Temporary contracts < 25th% percentile	-0.037***	-0.045***	-0.017	-0.026*
		(0.008)	(0.010)	(0.011)	(0.014)
	Temporary contracts > 75th% percentile	0.016**	0.013	0.014	0.012
		(0.008)	(0.010)	(0.011)	(0.014)
Productivity	Qualified employees < 25th% percentile	-0.043***	-0.026***	-0.050***	-0.025**
		(0.006)	(0.008)	(0.009)	(0.011)
	Qualified employees > 75th% percentile	0.001	0.001	0.003	0.001
		(0.006)	(0.008)	(0.009)	(0.011)
	Graduate employees < 25th% percentile	-0.076***	-0.070***	-0.064***	-0.056**
		(0.007)	(0.009)	(0.010)	(0.012)
	Graduate employees > 75th% percentile	0.029***	0.048***	0.025**	0.040***
		(0.008)	(0.010)	(0.010)	(0.013)
	Personnel development	0.329***	0.311***	0.332***	0.305***
		(0.007)	(0.010)	(0.010)	(0.014)

(to be continued)

(continued)

		Probit Regressions			
		all	comm.	all - 2007	comm. - 2007
Size effects	Number of employees	0.013**	0.017	0.004	0.022
		(0.006)	(0.014)	(0.008)	(0.019)
	$(\text{Number of employees})^2$	-0.000*	-0.001	-0.000	-0.001
		(0.000)	(0.001)	(0.000)	(0.002)
Reorganization	Reorganization of departments			0.069***	0.080***
				(0.010)	(0.014)
	Introduction of group work			0.090***	0.085***
				(0.016)	(0.021)
	Relocation of responsibility to the bottom			0.064***	0.074***
				(0.012)	(0.016)
Unionization indicators	Works council	0.060***	0.062***	0.062***	0.069***
		(0.008)	(0.010)	(0.011)	(0.014)
	Collective agreement	0.016**	0.018**	0.023***	0.032***
		(0.006)	(0.008)	(0.009)	(0.011)
	Company agreement	0.029***	0.019	0.023*	-0.004
		(0.010)	(0.014)	(0.014)	(0.017)
Log Likelihood		-12630.673	-6664.007	-6338.403	-3358.8257
Number of observations		31879	17669	16004	9004
Pseudo R^2		0.302	0.306	0.320	0.330

The regression further controls for industry, legal form, revenue, $(\text{revenue})^2$, commercial company (only in 1 and 3) and year. Robust standard errors in parentheses clustered for firm-ID. *** p<0.01, ** p<0.05, * p<0.1

Table 4.5: Determinants for the Use of Target Agreements

4.5.3 Effects of Introducing Target Agreements

In this section, we present the analysis of the performance effect after a firm implemented target agreements with employees between 2005 and 2007. The underlying sample of the regression analysis includes only firms without target agreements in 2005. Thus, the coefficient of the dummy for target agreements covers the within-firm difference in performance between those establishments that introduced target agreements between 2005 and 2007 and those that did not introduce it.

The first column of table 4.6 presents the first differencing OLS estimates of the log revenue. With any changes in other firm characteristics that may influence revenue growth, such as firm growth measured by the change in number of employees or the operation in an upcoming industry, establishments that introduced target agreements between 2005 and 2007 achieved revenue growth that was 5.7 percentage points higher than the revenue growth of firms without the practice (significant on the 1% level). This increase equals a revenue growth of 2.85 percent per year. Thus, the first-differencing approach delivers strong evidence for a performance effect on the organizational level. Column 2 in table 4.6 shows the result of the probit analysis using the dummy variable of the subjective evaluation of revenue growth as the dependent variable. In this specification, the introduction of target agreements for employees implies a significant positive probability increase of about 3.9 percent that company representatives expect revenue growth in the near future, which confirms the result. Organizations seem to profit from setting target for employees.

Some effects of the control variables are also worth mentioning. An increase in the turnover rate between the two years has a significantly negative effect on revenue, suggesting that a loss of human capital negatively influences firm‘s performance. Moreover, the percentage of qualified employees is positively correlated with revenue, which leads to the conclusion that stronger human capital may lead to revenue growth.

		First differencing - OLS	First differencing - Probit
		(1)	(2)
	Dependent variable:	Log revenue	Growth expectation (1/0)
Implementation	Target agreement (1/0)	0.057***	0.039*
		(0.015)	(0.021)
Tenure	Turnover rate	-0.168***	-0.031
		(0.059)	(0.024)
	Percent of empl. with temporary contracts	0.008	0.145**
		(0.092)	(0.063)
Productivity	Percent of qualified employees	0.074**	0.057*
		(0.032)	(0.030)
	Percent of graduate employees	0.043	0.059
		(0.062)	(0.061)
	Personnel development	0.011	0.016
		(0.015)	(0.019)
Size	Number of employees	1.280**	-0.001
		(0.612)	(0.141)
Constant		0.010	
		(0.006)	
Log Likelihood			-1897.0229
Number of observations		4813	3804
R^2 / Pseudo R^2		0.030	0.005

For all variables the differences $\Delta_{2007,2005}$ are considered. Additionally controlled for first differences of unionization variables. No controls are included for reorganizations, as the relevant questions were asked only in 2007. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 4.6: Performance Effects of Target Agreements

4.5.4 Separating the Effect from the Introduction of Performance Appraisals

One could argue that not the target agreements are the driving force of the performance effect here, but rather introduction of a formal performance appraisal system since targets are often the standards against which performance is subjectively appraised. Therefore, we further conducted a regression analysis in which we additionally add the dummy variable for the introduction of an appraisal system in organizations between 2005 and 2007. The variable is built in the same way as the dummy for introducing target agreements. Hence, the sample is further reduced to those firms that did not use performance appraisals and target agreements in 2005. Column (1) in Table 4.7 shows the result of the same regression as above when controlling for performance appraisal and the interaction between target agreements and performance appraisals. The interaction aims to account for the introduction of both practices between 2005 and 2007. The table shows that the effect of introducing target agreements becomes even larger (7.2%) and is highly significant, while the effect for introducing performance appraisals is notably smaller (4.1%) and significant at the 10 percent-level. The interaction term is not significant, suggesting that there is no additional utility for firms when both practices are implemented together. The results show that setting targets for employees in firms may have an even larger effect than the introduction of performance appraisal.

Argueing that targets for employees are even always linked to performance appraisals in organizations, we assumed for a further specification that target agreements cannot exist decoupled from performance evaluations. This approach required recoding of the dummy for performance evaluations to 1 in all cases in which target agreements are introduced. Only the two dummy variables, one for the introduction of performance appraisal alone and one for target agreements combined with an, in some cases, assumed performance appraisal are included in the first-differencing regression. Results are shown in column (2) of table 4.7. The introduction of performance appraisal alone does not seem to have a significant effect on performance; only when per-

	First differencing - OLS	
Dependent variable:	Log revenue	
	(1)	(2)
Target agreement	0.072***	0.057***
	(0.024)	(0.020)
Performance appraisal	0.041*	0.031
	(0.024)	(0.020)
Target agreements*Performance appraisal	-0.036	n.a.
	(0.041)	
Intercept	0.000	0.001
	(0.007)	(0.007)
Number of observations.	4120	4120
R^2	0.047	0.046

For all variables the differences $\Delta_{2007,2005}$ are considered. Additional control variables are the same as used in the regression in table 4.6. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 4.7: The (combined) Effects of Performance Appraisal and Target Agreements

formance appraisal is implemented together with target setting we observe a total revenue growth of 5.7 percent over those that have not introduced the practices. This result further strengthens the effect of target setting and suggests that setting goals as a performance benchmark may be a necessary condition for actually improving performance, also when linked to a formal evaluation system.

4.6 Conclusion

We investigated the determinants and productivity effects of using target agreements based on a representative firm level data set in Germany. This paper is the first to present evidence on which firms use target agreements for managing employees. As a formal target setting approach is costly to a firm, only those who may gain the highest benefits from it are likely to use it. We find evidence for a negative relationship between employee tenure and the use of target agreements. The necessity of incentivizing and directing high-

tenured employees decreases because of deferred compensation practices and their understanding of the firm‘s objectives. In addition, organizations with a higher percentage of qualified workers will have a greater loss if employees reduce their effort, so these organizations have a greater need for specific targets to make their employees exert more effort. Our analysis confirms this relationship. Target agreements for employees are also more likely to be used in establishments that have undergone job reorganizations. Employees working with new responsibilities or in newly introduced workgroups may need to be guided with targets more than other employees would, so guidance and alignment with company goals is more important. The firm size was only slightly positively related with using target agreements, although an economies of scale effect could be expected. Organizations with a higher level of labor union organization are more likely to use target agreements, which is also in contrast to our prediction, but this result suggests the need for further analysis of whether unionization in firms facilitates the introduction of HR-related activities in firms.

We find strong evidence for increased firm-level performance following the introduction of targets for employees. Our analysis shows that firms that introduced the practice between 2005 and 2007 saw higher revenue growth (by 5.7 %) than did firms that did not. By using a first differencing approach and controlling for structural changes like the change in the number of employees, we get close to a cause-effect analysis. Our result is robust when using managerial expectations of future revenue growth as dependent variable. Moreover, the effect of introducing combined target agreements and performance appraisal is larger than the effect for the introduction of only performance evaluations.

In addition to the strong evidence from the study, there are, of course, some shortcomings. Future investigations should consider alternative performance measures, such as firm profit or value added. Also we do not have precise information on the coverage of target agreements in firms, which would allow a more precise analysis about for which jobs or employees target setting is most profitable. However, missing this information makes our result even stronger because the result may be driven only by a share of

employees covered by the target agreements. Besides identifying a strong performance effect on the firm level, we cannot look at the mechanisms behind. Future researchers may be encouraged to collect comprehensive data sets from companies to understand these mechanisms and reactions by employees experiencing the introduction of target agreements. Also, methods of insider econometrics (like for example in the study of Ichniowski et al. (1997)) might be applied. Some recent studies have shown renewed interest in target setting and use rich data sets (Bol et al. (2010), Anderson et al. (2010)). While these rather focus on how targets are set and influenced by different parties, the employee reactions should receive further attention as well.

4.7 Appendix of Chapter 4

		Probit Regressions			
	Dependent variable:	Target Agreement (0/1)			
		(1)	(2)	(3)	(4)
		all	comm.	all - 2007	comm. - 2007
Tenure	Turnover rate	-0.008	-0.000	-0.021	-0.032
		(0.005)	(0.016)	(0.014)	(0.024)
	Share of empl. with temporary contracts	0.071***	0.118***	0.043*	0.061*
		(0.017)	(0.026)	(0.025)	(0.037)
Productivity	Percent of qualified employees	0.083***	0.059***	0.089***	0.054***
		(0.011)	(0.013)	(0.015)	(0.019)
	Percent of graduate employees	0.117***	0.188***	0.112***	0.158***
		(0.007)	(0.010)	(0.010)	(0.014)
	Personnel development	0.347***	0.333***	0.342***	0.317***
		(0.007)	(0.010)	(0.010)	(0.014)
Size effects	Number of employees	0.031***	0.055***	0.017*	0.060*
		(0.008)	(0.016)	(0.009)	(0.030)
	$\left(\text{Number of employees}\right)^2$	-0.001***	-0.003**	-0.000*	-0.003
		(0.000)	(0.001)	(0.000)	(0.003)

(to be continued)

(continued)

		Probit Regressions			
		all	comm.	all - 2007	comm. - 2007
Reorganization	Reorganization of departments			0.084***	0.091***
				(0.011)	(0.014)
	Relocation of responsibility			0.096***	0.093***
				(0.017)	(0.021)
	Introduction of group work			0.065***	0.076***
				(0.012)	(0.016)
Unionization indicators	Works council	0.094***	0.100***	0.082***	0.095***
		(0.008)	(0.011)	(0.010)	(0.014)
	Collective agreement	0.018***	0.022***	0.026***	0.035***
		(0.006)	(0.008)	(0.009)	(0.011)
	Company agreement	0.029***	0.018	0.023*	-0.005
		(0.010)	(0.014)	(0.014)	(0.017)
Log Likelihood		-12804.49	-6763.6692	-6380.8146	-3382.0898
Number of observations		31879	17669	16004	9004
Pseudo R^2		0.293	0.295	0.316	0.325

The regression further control for industry, legal form, revenue, $(\text{revenue})^2$, commercial company (only in 1 and 3) and year. Robust standard errors in parentheses clustered by firm. *** p<0.01, ** p<0.05, * p<0.1

Table 4.8: Determinants for the Use of Target Agreements - Continuous Variables

		Probit Regressions			
	Dependent variable:	Target Agreement (0/1)			
		(1)	(2)	(3)	(4)
		all	comm.	all - 2007	comm. - 2007
Tenure	Turnover 25th% percentile	-0.062***	-0.060***	-0.046***	-0.042***
		(0.008)	(0.009)	(0.011)	(0.013)
	Turnover 75th% percentile	-0.002	0.008	-0.001	-0.005
		(0.007)	(0.008)	(0.011)	(0.012)
	Temporary contracts 25th% percentile	-0.046***	-0.058***	-0.034***	-0.050***
		(0.008)	(0.009)	(0.012)	(0.013)
	Temporary contracts 75th% percentile	-0.003	-0.007	-0.010	-0.020
		(0.008)	(0.009)	(0.011)	(0.012)
Productivity	Qualified employees 25th% percentile	-0.038***	-0.021**	-0.035***	-0.018
		(0.008)	(0.009)	(0.011)	(0.012)
	Qualified employees 75th% percentile	0.002	0.001	0.005	0.002
		(0.007)	(0.009)	(0.010)	(0.012)
	Graduate employees 25th% percentile	-0.055***	-0.051***	-0.042***	-0.039***
		(0.008)	(0.010)	(0.012)	(0.014)
	Graduate employees 75th% percentile	0.052***	0.060***	0.050***	0.061***
		(0.009)	(0.011)	(0.012)	(0.014)
	Personnel development	0.315***	0.306***	0.317***	0.293***
		(0.009)	(0.011)	(0.012)	(0.016)

(to be continued)

(continued)

		Probit Regressions			
		all	comm.	all - 2007	comm. - 2007
Size effects	Number of employees	0.016*	0.037**	0.009	0.046
		(0.009)	(0.015)	(0.010)	(0.028)
	$(\text{Number of employees})^2$	-0.000*	-0.002	-0.000	-0.003
		(0.000)	(0.001)	(0.000)	(0.003)
Reorganization	Reorganization of departments			0.065***	0.075***
				(0.013)	(0.015)
	Relocation of responsibility			0.097***	0.088***
				(0.020)	(0.023)
	Introduction of group work			0.072***	0.073***
				(0.015)	(0.017)
Unionization indicators	Works council	0.055***	0.054***	0.061***	0.062***
		(0.009)	(0.011)	(0.013)	(0.015)
	Collective agreement	0.016**	0.019**	0.024***	0.033***
		(0.007)	(0.008)	(0.010)	(0.012)
	Company agreement	0.035***	0.018	0.029*	0.002
		(0.010)	(0.014)	(0.014)	(0.018)
Log Likelihood		-8510.7422	-5529.7957	-4227.19	-2750.9746
Number of observations		21766	21766	10882	7486
Pseudo R^2		0.3022	0.3058	0.3202	0.3293

The regression further controls for industry, legal form, revenue, $(\text{revenue})^2$, commercial company (only in 1 and 3) and year. Robust standard errors in parentheses clustered by firm. *** p<0.01, ** p<0.05, * p<0.1

Table 4.9: Determinants for the Use of Target Agreements - Moving Averages

Chapter 5

Do Employees Reciprocate to Intra-Firm Training? An Analysis of Absenteeism and Turnover[1]

5.1 Introduction

Organizational investments in intra-firm trainings are significant. In 2005, total costs of continuing vocational training (CVT) amounted to 1.6% of total labor costs in the EU-27, according to a recent study by the European Union[2] (Cedefop (2010)). Almost 70% of all German companies provided CVT and about 30% of the workforce in a given firm participated in training courses. In companies with more than 500 employees, even 90% provided intra-firm trainings.[3]

Companies invest in trainings because they want to enhance employee productivity by improving their knowledge and skills. When investments are

[1] This chapter is based upon Breuer and Kampkötter (2011a).

[2] Figures for training courses. Other forms of vocational training like on-the-job trainings or job rotation excluded.

[3] Furthermore, in large companies, training provision was formalized to a large extent, as more than 70% of these companies had a specific person/unit responsible for training, pursued training plans, prepared a training budget and measured participant satisfaction.

made, also the returns have to be evaluated, which is why former research has mainly focused on the effects of training participation on productivity, both on the individual and organizational level. On the individual level, an increased employee productivity should be reflected in higher wages (Bartel (1995), Barron et al. (1989), Lynch (1992), Barron et al. (1993), Veum (1995), Parent (1999)).[4] We argue, however, that besides improving the skills of employees, training might also lead to behavioral responses. When firms pay for general and specific training, employees may perceive this as a gift to which they may respond by higher effort or commitment. This logic is similar to the prediction of the efficiency wage literature stating that employers pay wages above the market-clearing wage to ensure higher efforts by employees (Akerlof (1982)).

In this paper, we want to discuss the application of the gift exchange framework to the intra-firm training context and empirically test whether employees show a behavioral reaction when being trained. By using personnel records of a large multinational firm, we observe variables other than wages that may reflect behavioral responses of employees: absence hours and the turnover behavior of employees. For both outcomes, we give a standard economic and a behavioral prediction which is then empirically tested.

We find that firm-sponsored training that aims at the accumulation of general skills is associated with lower employee absence. While both, human capital theory and behavioral arguments would predict a negative relation of training and absenteeism, they differ in their prediction of the effect duration. A reciprocal reaction might only imply a temporary decrease in absence, while increased opportunity costs from being trained should result in a more persistent decrease. Based on several analyses we find a temporary effect of training on absence behavior. Moreover, theories make different predictions for the analysis of turnover. While firm-sponsored training, for improving mainly general skills, should increase the turnover probability of employees because increased general human capital raises their market value according

[4]Organizational-level studies investigate the return of training investments based on firm-level surveys whereby productivity is measured by accounting figures like sales per year (Bartel (1994), Black and Lynch (1996), Barrett and O'Connell (2001)) or value added (Dearden et al. (2006), Zwick (2006)).

to standard theory (see among others Lincoln and Kalleberg (1996), Shaw et al. (1998), Batt (2002), Fairris (2004), and Haines et al. (2010)), receiving training may also be perceived as being valued by the firm and signal employer's confidence that employees stay with the firm. In contrast to the standard prediction, this could positively affect employees' loyalty to the firm. We indeed find that firm-sponsored training is negatively related to employee turnover. Moreover, the effect is largest for low-tenured employees.

We contribute to the training literature in several aspects. First, the available data for training research is mainly based on survey data. Definitions of training can be quite heterogenous for respondents and there might be a problem of respondents' difficulty to reliably report their training participation (Barron et al. (1997)). With personnel records we can overcome these shortcomings and thus follow the recommendation of Bartel (1995) to "focus on collecting more comprehensive data from companies".[5] Second, we extend the standard analyses on the effect of training participation on wages by focusing on the effect on non-monetary indicators: Absenteeism and turnover probability. For the analysis of absenteeism, the use of personnel records is especially useful because survey answers on absenteeism may underestimate the true absence time. Using absence and turnover as outcome variables, a behavioral perspective of training effects is offered which might partially explain the rather weak human capital effect of training that has been manifested in previous studies (e.g. Bartel (1995)).

The paper proceeds as follows. In section 5.2 and 5.3, the theoretical background is described and hypotheses are derived. Section 5.4 explains the data set and the methods for the empirical analysis. Descriptive statistics and regression results of the effect on sickness absence are presented in the first part of section 5.5, whereas the second part of the section analyzes the effect on turnover. Finally, section 5.6 concludes.

[5] Some studies have already based their analysis of training effects on company data sets so far. See Bartel (1995), Krueger and Rouse (1998), Breuer et al. (2010) and Fahr et al. (2010).

5.2 Training and Sickness Absence

In the recent literature, sickness absence has been typically referred to as unscheduled absence from the workplace (in contrast to scheduled absence like e.g. vacation). So far, a series of studies has used data on individual absence as proxy for employee effort or shirking behavior (see e.g. Winkelmann (1999), Riphahn (2004), Ichino and Riphahn (2005)). This is reasonable as e.g. Barmby et al. (1991) report that the majority of sickness absence is only a short-term phenomenon indicating a practice of self-certification of illness by employees in most cases. In the case of Germany, employees have sufficient degrees of freedom in self-certifying illness because a doctor's certificate is only needed starting from the third sickness day. Unmotivated employees may thus be induced to stay away from the workplace for at least two days, while wages are paid continuously by the employer.

But how may training participation affect the absence behavior of employees? In an economic setting, individuals are typically endowed with a stock of time which they can allocate to work or leisure. Participation in company trainings increases the human capital of employees and, hence, the opportunity costs of spending leisure time. Increased human capital can hence lead to a reduction of leisure time. "Voluntary" sickness absence, being a proxy for leisure time, should decrease. Since human capital accumulation implies a permanent productivity increase, employees who have been trained are expected to be permanently less absent than untrained employees.

However, another argument for predicting the effects of training on absenteeism is derived from the efficiency wage theory. Akerlof (1982) proposed that firms provide gifts for employees by paying a wage above the market-clearing or fair wage and employees respond by exerting effort in excess of the minimum work standard (positive reciprocity). The gift exchange logic is a widely accepted phenomenon and has frequently been confirmed in laboratory settings (see Fehr et al. (1993) for the first experimental study and Fehr and Gächter (2000) for a survey on reciprocity). Firm-sponsored training may be perceived as a gift by employees as it increased their labor market

value.[6] Employees may hence positively reciprocate to this gift by increasing their effort. In contrast to the human capital argument, reciprocal behavior may only be temporary. According to an experiment conducted by Gneezy and List (2006) who analyzed the duration of a reciprocity effect when above-market-clearing wages are paid, participants positively reciprocated only in the first few hours of the job while adapting to a lower effort level afterwards.

The gift exchange logic should especially apply when training investments target an increase in employees' general human capital because then the employees should generally benefit through higher wages as they become more valuable for other firms as well. Although standard human capital theory suggests that firms are not willing to pay for general training due to the risk of losing the return of investment when employees leave the firm, empirical evidence shows the opposite in practice. According to Barron et al. (1998), about 60% of the firms covered in the US firm level survey of the Employment Opportunity Pilot Program (EOPP) reported that the offered training invests in almost only general human capital. Furthermore, Becker (1962) noted that, in practice, a distinction between general and firm-specific human capital is not easy to make considering that "much on-the-job training is neither completely specific nor completely general".

Some related studies in the management literature already reported a positive relationship between training participation and job satisfaction, with satisfaction mediating the effect on absenteeism (see e.g. García (2005), Jones et al. (2009)). However, most of this research is based on survey data with low sample sizes and a cross-sectional data structure. Krueger and Rouse (1998) are, to the best of our knowledge, the first to analyze the direct relationship between training and absence hours based on personnel records of two firms. They find a negative effect of training participation on absence hours in the same week and conclude that employees liked to attend the courses. Here, we rather focus on the annual effect and especially investigate the persistence of the effect on absence behavior of employees.

Both, standard and behavioral theory, expect training participation to

[6] This argument has already mentioned in the conclusion of Barrett and O'Connell (2001).

have a negative effect on sickness absence of employees. However, different results may be expected for the duration of the effect. While the effect should be permanent in case of human capital accumulation, a behavioral response to training may only have a short-term effect. This leads to our hypotheses:

H1a: Intra-firm training participation has a sustainable negative effect on employee absenteeism (human capital perspective).

H1b: Intra-firm training participation has a temporary, negative effect on employee absenteeism (gift exchange perspective).

5.3 Training and Employee Turnover

The relationship between training participation and voluntary turnover probability has largely been discussed in the economic literature. In a fully competitive labour market, accumulation of general human capital should increase employee turnover when the increased productivity is not reflected in wages. Higher wage offers by competitive firms will hence increase the probability to leave the firm. As the accumulation of purely specific human capital mainly increases the employees' value for the current employer, wages might be raised by the current firm to reward higher productivity, but competitive firms will most likely not match the offer. Employees' turnover probability may hence be decreased.

However, in the observed firm the offered trainings are mainly of general nature. Trainings are assigned to six training categories which are Leadership & Communication, Language & Culture, Business Administration & Law, Research & Production, IT and Project & Process Management. As the notations suggest, most of the trainings offered under these categories teach general contents whereas some categories such as Business Administration & Law may build up more transferable knowledge than for example IT trainings which may in some cases focus on a specific software tool. According to the standard rationale, the firm's investment in general skills should increase employees' turnover probabilities. But the opposite might be true when employees perceive firm-sponsored trainings as a gift from the firm. Receiving training may also help employees feel as an inherent part of the firm or

signal future career prospects within the organization. A positive reciprocal reaction of employees might result, so that turnover probability can even be reduced.

Former research already raised that the effects of training in practice may not always be consistent with human capital theory predictions which mainly focuses on the relation between training and wages (Levine (1993)). Also, labor market frictions (Acemoglu and Pischke (1999)) may limit the mobility of employees and hence offers an explanation why firms invest in general training at all. Some management studies emphasized the impact of intra-firm training on job satisfaction or commitment (see for example Lee and Bruvold (2003), García (2005), Jones et al. (2009)) as predictors of turnover intentions. Furthermore, several firm-level studies have investigated the human capital prediction of training and employee turnover (see for example Lincoln and Kalleberg (1996), Shaw et al. (1998), Batt (2002), Fairris (2004), Haines et al. (2010)). While Haines et al. (2010) have shown a positive correlation between intra-firm training and individual turnover rates, Lincoln and Kalleberg (1996), Shaw et al. (1998), and Batt (2002) did not find any significant relationship. Based on survey data from about 500 establishments, Fairris (2004) presented evidence on a slightly negative relationship. Besides the limited reliability of survey data, most of the studies are based only on cross-sectional surveys and therefore no temporal dimension can be included in the analysis. Individual-level analyses which are typically based on micro-survey data sets mainly look at the employability of workers on the labor market (see for example Picchio and van Ours (2011)) instead of analyzing individual turnover probability. According to the discussion above, we formulate two competing hypotheses:

H2a: Participation in intra-firm training has a positive effect on the turnover probability of employees (human capital perspective).

H2b: Participation in intra-firm training has a negative effect on the turnover probability of employees (gift exchange/commitment perspective)

5.4 Data and Empirical Approach

In the following, we will introduce the data set and the empirical strategy for the analysis. Subsequently, we will give an overview of the training participation and the training policy in the firm.

5.4.1 Data and Measurement

We investigate personnel records from a large, multinational company with headquarter in Germany for the years $2006 - 2008$.[7] The records comprise annual information on about $15,000$ German full-time, permanent employees resulting in a total of about $46,300$ employee-year-observations. The panel data set covers information about the training participation of employees, the number of attended trainings and the training content. According to the company's training policy, there are different reasons for training participation. We exclude mandatory trainings which typically inform about new legal requirements and have to be completed by all employees in the company. Only trainings that are designed to increase human capital are considered for the analysis. Employees can either choose to participate in these trainings or they are selected by their direct supervisor. Note that we observe two types of employees: managerial and non-managerial employees. Managerial employees typically have an academic background, while non-managerial employees have completed a vocational education.[8] As managerial employees in this firm have flexible working time arrangements, absence hours for these employees are not officially recorded. Hence, we only observe individual absence hours for all non-managerial employees. We look at absence hours due to illness reasons, which include absence without a doctor's certificate. Absent time due to holidays or other compensated absence (e.g. participation in trainings, works councils, etc.) is excluded, so that only those absence hours are considered which occur due to illness reasons. Note that we ex-

[7] Due to confidentiality reasons, the data sheets and the company name had to be anonymized.

[8] In Germany, non-managerial workers are normally tariff workers whose working contracts underlie the tariff commitment of the union with the respective company or the organization of employers.

clude employees from the analysis with annual absence hours lying above the 99th percentile of the distribution. As these employees are absent from work up to each working day in a year, considering employees on long-term sick leave may lead to biased coefficients.

Employee turnover is tracked by a dummy variable turnover with the value 1 if an employee is not employed by the firm in the following year and 0 otherwise. Hence, turnover can be measured for the years 2006 and 2007. We cannot explicitly differentiate between voluntary and involuntary turnover, but discussions with company representatives revealed that the share of involuntarily leaves is very low. According to company records in 2007, only 1.4% of employees who left the firm had to leave the firm involuntarily (15 out of 1,065 employees). Turnover is hence almost completely voluntary in the firm. Additionally, Germany has very rigid labor market laws protecting employees from dismissal, i.e. involuntary turnover is less often observed than in other countries. Note that we excluded employees older than 60 years of age in the turnover regressions to avoid potential effects of early retirement programs that are quite common in larger German companies.

Moreover, besides demographic information on age, years of firm tenure, and gender, we observe job-related information on annual salary, employee status, hierarchical level and business unit. There are 23 levels in the observed company. We measure promotions by a move to the next higher level. *Employee status* comprises four groups: Non-managerial employees (levels 2 to 12) and three different groups of managerial employees. These cover junior managers (levels 13 to 15), which include the typical entry positions for university graduates, senior managers (levels 16 to 17) and senior executives (levels 18 to 24). Table 5.1 shows that about two third of all employees in this firm are non-exempt employees, about 30% are working as junior or senior managers and about 3% are senior executives.

Average firm tenure (age) is 21 (42) years for non-managerial employees while it ranges from 15 to 19 (42 to 49) for managerial employees. We further observe three main subdivisions: The holding, the service units, and the operational/industrial units. About 4% of all employees work in the

Employee status	2006		2007		2008	
	Freq.	Percent	Freq.	Percent	Freq.	Percent
Non-managerial	9,879	67.84	10,517	67.32	10,254	65.25
Junior manager	2,559	17.57	2,709	17.33	2,980	18.95
Senior manager	1,779	12.22	1,988	12.72	2,042	12.99
Senior executive	345	2.37	411	2.63	441	2.81
Total	14,660	100.00	15,744	100.00	15,839	100.00

Table 5.1: Distribution of Employees by Employee Status and Year

holding, 22% in the service units and about 74% in the operational units.

5.4.2 Empirical Strategy and Identification

Our first hypothesis relates to the effect of training participation on absenteeism of employees. Note that only non-managerial employees can be considered for this analyses. We estimate the following equation to analyze if the participation in training in period t and $t-1$ has a positive effect on absenteeism of an employee i in year t:

$$a_{it} = \beta_0 + \beta_1 p_{it} + \beta_2 p_{it-1} + + c'_{it}\gamma + \upsilon_t + \alpha_i + \varepsilon_{it}$$

a_{it} indicates the absent hours due to illness by non-managerial employees. p_{it} and $p_{i(t-1)}$ are dummy variables taking the value 1 if person i participated in a training in year t or $t-1$ and 0 for no training participation. c'_{it} is a row vector of controls and covers several individual and job-specific variables such as age, tenure, logarithm of base salary, required working hours, level dummies and business unit dummies. The business unit fixed effects account for structural and cultural differences between units (such as peer pressure or demand shocks). To proxy human capital and job requirements, we include level dummies in the analysis. These covariates controls for differences in absence behavior that are correlated with the hierarchical position of an employee and for which different training policies might be applied. To rule out that training participation negatively affects absenteeism because of the necessity to finish ongoing tasks that had to be neglected during training

time, we additionally control for overtime hours of an employee (included in c'_{it}). β_0 is the intercept and β_1, β_2 and β_j are the estimated coefficients, υ_t are year dummies and α_i describe individual fixed effects to control for unobserved, time-constant heterogeneity, like individual ability or a general attitude towards absenteeism. ε_{it} denotes the time-dependent error term. According to our hypothesis, we expect a negative relationship between absent hours and training participation in year t. If the lagged effect shows a significant negative effect on absenteeism, a permanent human capital accumulation as predicted by the human capital theory seems prevalent. In a further specification we also include the interaction term between p_{it} and p_{it-1} to analyze if the repeated training participation has a stronger effect on absenteeism which would be in line with predictions of the human capital theory. We exclude employees from our sample who were promoted, newly hired or moved between subdivisions because they are expected to have different training policies than other employees.

Fixed effects least squares estimators are used to estimate the model for absence hours.[9] We aim at identifying the causal effect of training participation on absence behavior. Several conditions have to be met to ensure that these parameters are unbiased. Training participation has to be exogenous so that there is no selection of specific employee groups into training. One prominent argument (also discussed by Bartel (1995)) is that higher-skilled employees are more likely to be trained by firms because of a higher expected return on investment. Assuming that relevant unobserved characteristics for selection into training, such as ability, do not vary over time, this selection effect is controlled for when applying individual fixed effects in the regressions. But the individual fixed effects do not eliminate the bias resulting from time-varying heterogeneity. As employees might reveal their productivity over time, this might still be included in the time-dependent error term of the regression. We try to eliminate this bias in a further specification of the model by including a productivity measure into the regression. Productivity is proxied by the percentile of individual's overtime hours and

[9] We also computed tobit estimates with 0 as the lower limit yielding the same results. For interpretation purposes, the least squares estimates are reported.

the percentile of the individual's base salary in the respective peer group in year $t-1$ (Bartel (1995) uses a similar ability proxy). A peer group is defined as a unique combination of year, strategic business unit (subgroups of business unit) and salary level, with an average of 86 individuals.[10] If the coefficient of training participation on absence hours does not substantially vary in the specification with and without the productivity measures, endogenous selection of more productive employees into training is unlikely to bias the estimates. One might argue that the use of the salary percentile as productivity measure is inappropriate because a position in the upper part of the salary band may only compensate an employee for not being considered in future promotions or simply reflect seniority wages (Lazear (1981)). In a companion paper (Breuer and Kampkötter (2011b)), we tested the reliability of this measure by regressing the promotion probability on several salary percentile dummies. An increasing salary percentile seems is significantly increases the promotion probability which supports the use of the percentile as productivity measure.

We conduct a further robustness checka and estimate separate fixed effects regressions both for high and less productive employees. If the coefficient of training participation is robust for the subsample of employees located at the bottom of the overtime or salary distribution, selection of productive employees will be less likely to bias the parameters. In another robustness check, we further address the selection problem by applying a kernel matching approach to estimate the average treatment effect on the treated (ATT). This approach attempts to eliminate the endogenous selection problem by finding an optimal control group which does not systematically differ to the treatment group. By comparing the outcome of treatment and control group, the ATT can be estimated. The control group is computed based on propensity scores (a kind of probabilities) of training participation which are derived based on observable individual and job covariates. For estimating the propensity score we, for instance, take into account whether an employee has participated in training the year before the period of interest. In addition,

[10]Note that we exclude groups with less than three observations for the computation of the percentile.

the productivity proxies from above are included as predictors of training participation. Finally, absent hours of the treated group (training participation) are then compared to the untreated group (no training participation) in the same year. A detailed explanation of the matching procedure is included in the robustness check section of this paper.

In the analysis of turnover (hypothesis 2), we use a probit regression function with the turnover dummy as the dependent variable. This analysis is conducted for all employees in the sample.[11] It is argued that training participation in t may have an impact on the probability to leave the firm in $t+1$ so that the equation

$$P(Turnover_{i(t+1)} = 1|p_{it}, X_{it}) = G(\lambda + \delta p_{it} + \eta X_{it})$$

describes our model. In the probit model, G is the the standard normal cumulative distribution function. The effect of training on the probability of turnover is indicated by the estimated coefficient δ, whereas λ is the intercept and X_{it} covers several control variables that may influence the probability of turnover. These are firm tenure, age, gender, fixed wage, level dummies, business unit and year fixed effects. In addition, we include an ability proxy in this model because more able employees are expected to have better outside options for being employed at another firm. This measure is the above-defined percentile position of an individual employee in the salary distribution of the direct peer group. In contrast to the percentile in the overtime distribution, the salary percentile can be calculated for all, managerial and non-managerial, employees. The definition of the peer group is identical to the one above. By including this measure we can partially rule out that the coefficient of training is positively biased through low-ability employees receiving less training and being more likely to leave the company. In one further specification we additionally include an explanatory dummy variable indicating if an employee participated in training in $t-1$. Like in the analysis of absenteeism, we exclude employees from the sample that are

[11] As we only have three years of data, we are not able to run harzard rate models to predict turnover probabilities. Hence, we follow other studies using a linear probability model (see for example Krueger and Rouse (1998)).

promoted, newly hired or just moved between subdivisions. Overall, we can make use of about 30,000 observations (from 2006 and 2007) for the analysis of turnover.

We also conduct robustness checks to address a potential selection bias of the estimated parameters. We use two further productivity measures as controls in additional regressions. These are the overtime percentile for non-managerial employees and individual's bonus percentile in the relevant peer group for managerial employees.[12] To rule out the possibility that an effect is only driven by the more productive employees, we further conduct separate regressions for the above and below-median group of individuals with respect to salary, overtime hours and bonus payments. Finally, a matching approach is also applied for the analysis of turnover.

Table 5.2 provides descriptive statistics for our dependent and independent variables. The mean of absence hours is 73.5 hours (about 9.2 days) with a median of 36 hours (=4.5 days), the average turnover rate of the three years is 6.5%.

5.4.3 Training Participation in the Firm

In the following, we aim to descrobe how training is applied in the observed firm. Typically, classroom trainings are offered to a group of employees and are instructed by an internal or external trainer. In our data set, we observe that about one third of all employees attained at least one training in a respective year (15,076 training-year observations during the time span 2006-2008). This indicates that classroom trainings are a widely used personnel development instrument in the analyzed company. Overall, training participation decreased from 36% in 2006 to 31% in 2008. The rate of training participation may depend on the economic situation of a company as in growth periods employees may have fewer days left to participate in training programs. There is some indication for this relationship as the observed firm increased net sales significantly between 2006 and 2008. Figure 5.1 shows the

[12] The reliability of the bonus percentile is also tested in our companion paper (Breuer and Kampkötter (2011b)).

Descriptive Statistics	Obs.	Mean	Std. dev.	Min	Max
Dependent variables					
Absent hours (non-managerial)	30,650	73.526	108.362	0	885
Turnover (0/1) (all employees)	30,116	0.065		0	1
Training indicators					
Training participation t (0/1)	45,874	0.329		0	1
Training participation $t-1$ (0/1)	27,776	0.357		0	1
Number of trainings t	45,874	0.636	1.201	0	13
Individual pay & hierarchical level					
Fixed salary (in EUR)	45,353	(conf.)$^{\alpha}$	27,584	(conf.)$^{\alpha}$	(conf.)$^{\alpha}$
Bonus payment (in EUR)	14,835	(conf.)$^{\alpha}$	15,130	(conf.)$^{\alpha}$	(conf.)$^{\alpha}$
Non-managerial employee (0/1)	45,874	0.668		0	1
Junior managers (0/1)	45,874	0.180		0	1
Senior Manager (0/1)	45,874	0.127		0	1
Senior Executive (0/1)	45,874	0.026		0	1
Working time (non-managerial)					
Overtime hours (unpaid)	30,650	10.898	39.138	0	717
Individual characteristics					
Tenure	45,874	19.721	10.021	0	49
Age	45,874	42.937	8.906	19	64
Female (0/1)	45,874	0.199		0	1
Individual job moves					
Promotion (0/1)	27,801	0.125		0	1
Newly hired (0/1)	45,874	0.015		0	1
Division move (0/1)	27,868	0.020		0	1

$^{\alpha}$ confidential

Table 5.2: Descriptive Statistics

average training participation (in at least one training) by employee status and year (promoted, newly hired and move employees excluded). Junior managers show the highest rates with a three-year average of about 52%. It shows

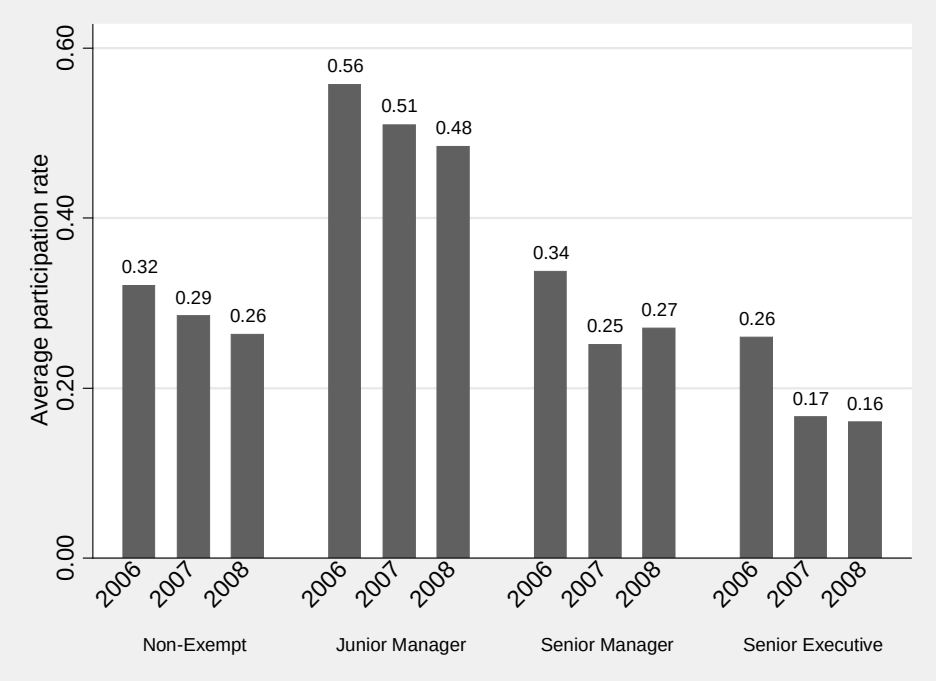

Figure 5.1: Average Training Participation by Employee Status and Year

that the focus of personnel development instruments like classroom trainings is on young managerial workers which are supposed to acquire human capital to become the future leaders of the company. Besides participation rates, the data set covers information on training hours and the number of attended trainings. On average, a training lasts 19.5 hours in this company, which is equivalent to about 2.4 training days. Table 5.3 shows the average number of trainings per employee and year for trained employees.

Junior managers are the mostly trained employee group and also receive the highest number of trainings (two trainings per year on average). As mentioned above, trainings are assigned to six categories which are "Business Administration and Law" (with overall participation rate 5.2%), "Research and Production" (4.2%), "IT" (8.1%), "Leadership and Communica-

	Number of trainings		
Employee status	2006	2007	2008
Non-managerial	1.95	1.82	1.79
Junior manager	2.50	2.02	2.11
Senior manager	1.77	1.63	1.70
Senior executive	1.60	1.52	1.51
Total	2.07	1.85	1.88

Table 5.3: Average Number of Classroom Trainings by Employee Status and Year

tion" (8.5%), "Project and Process Management" (3.6%) and "Language and Culture" trainings (8.5%). Participation rates are the highest in "Leadership and Communication" and "Language and Culture" trainings which is mainly driven by junior managers, too.

5.5 Results

5.5.1 Effects of Training on Worker Absence

Prior to the interpretation of the regression results, we provide some descriptive evidence on the relationship between training and individual absenteeism. Table 5.4 shows a positive trend in absence hours over the years with much lower absence hours for non-managerial employees that participate in training compared to those who do not participate. Employees without training participation are absent from work about 23.8 hours more than trained employees (t-test, p=0.000).

Average absence hours	2006	2007	2008	Total
Training Participation	52.3	56.9	61.5	56.7
No Training Participation	76.2	79.7	85.0	80.5

Table 5.4: Absence Hours by Training Participation

The results of the fixed effects regression are shown in table 5.5. Controlling for unobserved, time-constant individual heterogeneity, column (1)

shows a negative and significant coefficient of training participation indicating a decrease of 7.9 absence hours. When additionally including training participation in $t-1$ (column 2), the table shows that participation in at least one training leads to a substantial decrease of 10.7 absence hours in the same year, which is equal to about 1.3 working days. Based on the average annual salary paid to non-managerial workers, this yields a 'recovered value' of about EUR 250 (equal to USD 350, exchange rate = 1.3994 EUR/USD on 31/12/2008) for the participation in training per employee. However, the effect of training participation in the year $t-1$ is not statistically different from zero indicating that no sustainable effect on sickness absence behavior is observed. This supports the hypothesis of a reciprocal reaction.

Additionally including the interaction term between training in t and training in $t-1$ in the model shows that the negative effect on absenteeism for trained employees in the current year is even stronger when an employee was not trained in $t-1$ (see column (3)). The results suggest that employees adjust there absence behavior, dependent on training participation, only in the current year. The short-term effect rather leads to the conclusion that employees temporarily reciprocate to firm-sponsored trainings.

Columns (4) to (6) show that the estimates are robust when the productivity proxies of the year $t-1$ are included. Assuming that the measures are reliable proxies for time-varying productivity, the estimated effect of training is hence not biased by an endogenous selection of more productive employees into training. Interestingly, unpaid overtime hours show a negative coefficient, but the effect is mostly not statistically significant. By further assuming that unpaid overtime is a proxy for work motivation, using it as a control variable can mitigate the problem that the effect of training on absence behavior is driven by more motivated or highly engaged employees who are also more likely to participate in trainings and are less absent.

Dependent variable:	Absence hours$_t$					
	(1)	(2)	(3)	(4)	(5)	(6)
Training$_t$	-7.882***	-10.716***	-14.755***	-11.427***	-10.256***	-14.394***
	(1.934)	(3.517)	(4.817)	(3.288)	(3.582)	(4.895)
Training$_{t-1}$		2.169	-1.165		2.457	-0.968
		(3.383)	(4.342)		(3.455)	(4.428)
Training$_t$*Training$_{t-1}$			8.648			8.919
			(5.646)			(5.746)
Overtime hours$_t$	-0.083**	-0.053	-0.050	-0.059	-0.062	-0.059
	(0.042)	(0.057)	(0.057)	(0.058)	(0.057)	(0.057)
Tenure^2	-0.037	0.253	0.249	0.233	0.233	0.228
	(0.070)	(0.155)	(0.154)	(0.156)	(0.156)	(0.156)
Age^2	0.041	-0.321*	-0.317*	-0.312*	-0.312*	-0.307*
	(0.077)	(0.172)	(0.172)	(0.175)	(0.175)	(0.175)
Individual fixed effects	yes	yes	yes	yes	yes	yes
Productivity proxies$_{t-1}$				yes	yes	yes
Observations	26,876	15,266	15,266	14,979	14,979	14,979
R^2 within	0.01	0.01	0.01	0.01	0.01	0.01
Number of groups	11,548	8,896	8,896	8,763	8,763	8,763

Additional control variables include ln salary, required working hours, business unit, salary level and year dummies. Including lagged training participation leads to an decrease in observations Column (4) - (6): percentile position in $t-1$ included which further reduces the sample size Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 5.5: Fixed Effects Regression Results for Absence Hours

5.5.2 Robustness Tests for Predicting Employees' Absence

As mentioned above, we re-estimate our model separately for high and low productive employees. Employees with overtime hours and salary above the peer group median in period $t-1$ are assumed to be more productive, employees below the median to be less productive. The results of the four subsamples are reported in table 5.6. In all specifications, the negative effect can be confirmed. Employees with above and below-median salary or overtime hours show a decrease in absence hours when they are trained in the same year, with an even larger effect for the below-median groups. But the coefficient difference between the groups is only significant (on the 5% level) when comparing the above and below median groups of the salary percentile. Hence, especially less productive employees may perceive training as a gift from the company.

Dependent variable:	Sickness Absence hours$_t$			
Grouping variable	Salary percentile		Overtime percentile	
	$>$ median	$\leq$ median	$>$ median	$\leq$ median
Training$_t$	-5.502**	-17.859***	-7.714**	-9.792**
	(2.417)	(6.167)	(3.328)	(4.203)
Overtime hours$_t$	-0.204***	-0.198***	-0.212***	-0.516
	(0.064)	(0.075)	(0.045)	(0.343)
Observations	21,564	5,312	16,946	9,930
R^2 within	0.01	0.02	0.02	0.005
Number of groups	11,522	3,567	11,520	6,411

Control variables are the same as included in table 5.5.

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

Table 5.6: Fixed Effects Regression - Effect of Training Participation on Absence Hours for Subgroups

We additionally conduct a matching approach which has become popular as a possible solution to the selection problem. The following paragraph explains this approach in detail. In general, matching allows to non-parametrically test the average effect of a binary variable on an outcome variable (for a survey see Caliendo and Kopeinig (2008)). Assuming that

there are two potential outcomes Y_i^1 (employee i receives treatment = training, $T = 1$) and Y_i^0 (employee i receives no training, $T = 0$), the treatment effect is described by the difference in potential outcomes $\tau_i = Y_i^1 - Y_i^0$. As we will never observe both outcomes for one person at the same time, an ideal treatment effect could be derived by comparing the outcome of treated and untreated individuals under the condition that both groups are intended to receive training. An ideal average treatment effect on the treated (ATT) would be described as:

$$\tau_{ATT} = E(Y^1|T = 1) - E(Y^0|T = 1)$$

The second term is the hypothetical term; it is the outcome of individuals which are meant to receive training but are not treated. In the sample, we only observe $E(Y^0|T = 0)$ which is typically not equal to $E(Y^0|T = 1)$ because the outcome of treated individuals may systematically differ from untreated individuals. Only comparing the outcomes of the two subsamples could therefore imply a selection bias. But comparing the outcome of the treated with the outcome of an *adequate* control group can be done by applying a matching estimator. The basic idea is to find, for each treated individual i with characteristics X_i, a control individual with the same characteristics and then compare their outcomes. One established method for determining the control group is the method of propensity score matching (first discussed by Rosenbaum and Rubin (1983)). The propensity score is the probability of receiving treatment (participation in training) given a vector of observable covariates X_i. Individuals with the same propensity scores are divided into two groups, the treated and the untreated. Average values of the variables that are used to predict the propensity score should be balanced for both groups. The assignment to training is then supposed to be random for individuals having the same propensity score. Two conditions have to hold when conducting propensity score matching: the conditional independence assumption (CIA) (Lechner (1999)) stating that conditional on the propensity score $P(X_i)$, the counterfactual outcome is independent of treatment: $Y^0 \perp T|P(X_i)$. This is a strong assumption implying that all

variables that influence treatment assignment and outcome simultaneously are to be observed by the researcher. Variables included in the matching procedure therefore have to be carefully selected (Smith and Todd (2005)). The second condition to hold is 'overlap' and requires $Pr(T = 1|P(X_i)) < 1$. It means that the probability of not participating in training has to be positive for every propensity score $P(X_i)$. When these assumptions hold, the ATT can be identified by:

$$\tau_{ATT}^{Matching} = E(Y^1|P(X_i), T = 1) - E(E(Y^0|P(X_i), T = 0)|T = 1)$$

We estimate the propensity score of participating in trainings based on a probit model. Two propensity score estimations are computed here, one for training participation in the same year and one for participation in $t-1$. The observable covariates that are used to predict training participation have been carefully selected based on economic rationales and described in the following.[13] In addition, we checked econometric indicators such as significance or the pseudo-R^2 (which should be as high as possible) to define the final probit specification for determining the propensity score regression (for further discussion of estimating propensity scores see e.g. Heckman et al. (1998)).

The observable covariates in the probit regression of training participation are gender, age, tenure, the logarithm of base salary, overtime hours, business unit and year dummies. To take into consideration that higher-positioned employees have different training habits due to time constraints, we include level dummies to predict training participation. According to the argument that more able employees are more likely to be selected into training, we include the salary percentile and the overtime percentile in the peer group's distribution as ability proxy. Furthermore, we include training participation in $t-1$ as a predictor of training in t because trained employees are more likely to be trained also in the subsequent year. In addition, as absence hours is the defined outcome variable, we also add absence hours in $t-1$ as predictor variable to the probit regression. If employees are absent

[13]Insights from others studies with regard to the determinants of individual training participation have been considered (see for example Bartel (1995)).

due to sickness in $t-1$, firms may not want to invest in their human capital in t because they are regarded as unproductive or employers anticipate that these employees will also be absent in the future implying that the newly accumulated skills cannot be applied properly. One further determinant is the number of trainings the direct supervisor has participated in. When the supervisor regularly participates in trainings, this may induce a training culture with spill-over effects on the employees. It may also reflect an increased requirement of human capital accumulation in a given department. We excluded employees that were promoted, newly hired or moved subdivisions from the analyses. The results of the probit regressions are shown in table 5.13 in the appendix. Note that overall significance levels of coefficients are the same in both regressions. The results show that training participation in $t-1$ is highly correlated with training participation in t. Also, women are more likely to participate which might be due to career breaks of women when giving birth to children leading to a need to catch-up with men afterwards (Fitzenberger and Muehler (2011)). As predicted by human capital theory, older workers are less likely to participate in training possibly because of a shorter amortization period for the training investment. However, we see that employees located higher in the salary distribution are less likely to participate in training.

The distribution of propensity scores is illustrated in figures 5.3 and 5.4 in the appendix and indicates that the propensity score distribution of the true participants is slightly skewed to the right. This suggests that the true training participants are, on average, more likely to participate in trainings. Nevertheless, participant's propensity score overlaps the region of propensity scores of non-participants completely. Hence, the overlap condition is fulfilled. We conducted several quality checks of the propensity score matching which are shown in columns (1) and (2) of table 5.14 in the appendix.[14] By computing t-tests in order to compare the means of covariates between participants and non-participants before and after matching, we see that although there were significant mean differences on the 5% level for 20 to 25 covariates before, there are no significant differences (on the 5% level) after

[14]For a discussion of quality checks see Caliendo and Kopeinig (2008).

the matching procedure. On the 10% level, there is a significant difference in the mean for only one covariate. The affected variable is a dummy covariate for one hierarchical level. Also, the standardized bias is reduced substantially after the matching procedure (see table 5.14). Lastly, we compared the Pseudo R^2 when predicting training participation in the probit regression for the samples before and after matching . Pseudo R^2 is about zero when using the matched sample which confirms that propensity score matching was successful.

In the next step, we estimate the ATT from the equation above by using a kernel matching algorithm in order to be able to run bootstrapping for inferences.[15] The ATT for both specifications are shown in table 5.7.

Outcome variable: Absent hours$_t$	Obs	Difference	Bootstrap S.E.	P>|z|
ATT of Training$_t$	14350	-13.242***	1.989	0.000
ATT of Training$_{t-1}$	6529	-2.218	3.964	0.723

Standard errors are based on bootstrapping with 200 replications.
*** p<0.01, ** p<0.05, * p<0.1.

Table 5.7: Propensity Score Matching Result - Training in t and t-1 on Absent Hours

For the estimation of standard errors, bootstrapping with 200 replications was conducted. When comparing the absent hours of the treated sample (those who have participated in training) with the absent hours of the matched sample, we see that employees who participated in a training in t are significantly less absent (13.2 hours) than their matched sample which confirms the results from above. The ATT of the group that were treated in $t-1$ is also negative, but far from statistical significance. Employees that were trained in the previous year do not show a different absence behavior today than employees who were not trained in $t-1$. Overall, the matching procedure confirms our results from above that there is only a temporary effect of training participation on absence behavior.

[15] Precisely, we conducted the Epanechnikov Kernel with a bandwith of 0.06. Additionally, other matching algorithms (e.g. Nearest-Neighbour matching) were computed and the results were confirmed in all specifications.

	Turnover rate t+1	
	(1)	(2)
No Training (=0) in t / t-1	8.2%	6.8%
Training (=1) in t / t-1	2.3%	3.0%
Difference in turnover[a]	5.8%***	3.8%***

[a] Tested with t-test (*** denotes significance on 1%-level). Column (1) reports figures for training or no training in t and column (2) for t-1

Table 5.8: Turnover Rates by Training Participation

5.5.3 Effects of Training on Turnover Probability

To analyze the relationship between training participation and turnover probability, we first present descriptive statistics for differences in turnover rates. As shown in table 5.8, turnover rates are significantly lower for employees that have been trained in the actual or previous period. We see in column (1) that the turnover rate for the group of employees that have not been trained (8.2%) in t is three times larger compared to the trained group (2.3%). The difference in turnover is smaller for the groups separated by training participation in $t-1$, but still significant (see column (2)).

We further test the impact of training participation based on a probit model with the binary variable 'turnover in t+1' as dependent variable. Table 5.9 reports marginal effects for all specifications. Column (1) shows that employees that were trained in period t are significantly less likely of leaving the firm voluntarily in $t+1$. With -5.0% this effect is substantial showing a strong correlation between training and future turnover of an employee. Controlling for training participation in $t-1$ in column (2), the effect is confirmed and reveals that employees who receive training in the previous year are also less likely to leave the firm voluntarily. As only a mixture of general and specific trainings are offered, this effect is contrary to the one expected by human capital theory and rather supports the hypothesis of increased attachment of employees after training participation. The result stands in contrast to Krueger and Rouse (1998) who report a small, but positive relationship between training participation and quitting behavior.

Dependent variable:	Turnover$_{t+1}$ (0/1)	
	(1)	(2)
Training participation$_t$	-0.050***	-0.026***
	(0.002)	(0.003)
Training participation$_{t-1}$		-0.029***
		(0.003)
Percentile Salary$_t$	0.025***	0.032***
	(0.006)	(0.008)
Ln(Salary)$_t$	-0.250***	-0.263***
	(0.010)	(0.016)
Tenure$_t$	-0.001***	-0.000
	(0.000)	(0.000)
Age$_t$	-0.000	-0.001*
	(0.000)	(0.000)
Female$_t$	0.032***	0.022***
	(0.004)	(0.005)
Observations	27,018	11,062
Pseudo R^2	0.2193	0.231
Log Likelihood	-5044.037	-1750.955

Marginal effects reported. Additional control variables include salary level, subdivision and year. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 5.9: Probit Regression Results for Employee Turnover

Some control variables also show interesting insights. The more productive employees (proxied by the salary percentile) are more likely to leave the firm which confirms that they may have better outside options on the labor market. Furthermore, turnover probability decreases with the salary paid to employees. This is in line with the prediction that higher salaries tie employees to the firm, holding other job-related factors like tenure and salary level constant. Moreover, women show a significant higher turnover probability which might be driven by females leaving the labour market to take over family responsibilities.

However, the effect of training participation on the loyalty of employees may depend on individual's firm tenure. First, mobility or turnover probabilities are generally higher for low-tenured employees because they may

have accumulated less specific human capital as for example through networks or on-the-job learning effects. Hence, there is more room to increase their attachment to the firm compared to higher-tenured employees. Second, higher-tenured employees already might have received several firm-sponsored trainings during their time with the firm, so that the marginal effect of training on their behavior is decreased the longer the employee has worked for the firm. To check this, we computed the same probit regression from above for eight tenure categories: 0-4 years, 5-9 years, 10-14 years, 15-19 years, 20-24 years, 25-29 years, 30-34 years and more than 34 years of tenure. Figure 5.2 shows the predicted turnover probabilities (at the means of covariates) for the eight tenure groups in dependence of training participation. We see that

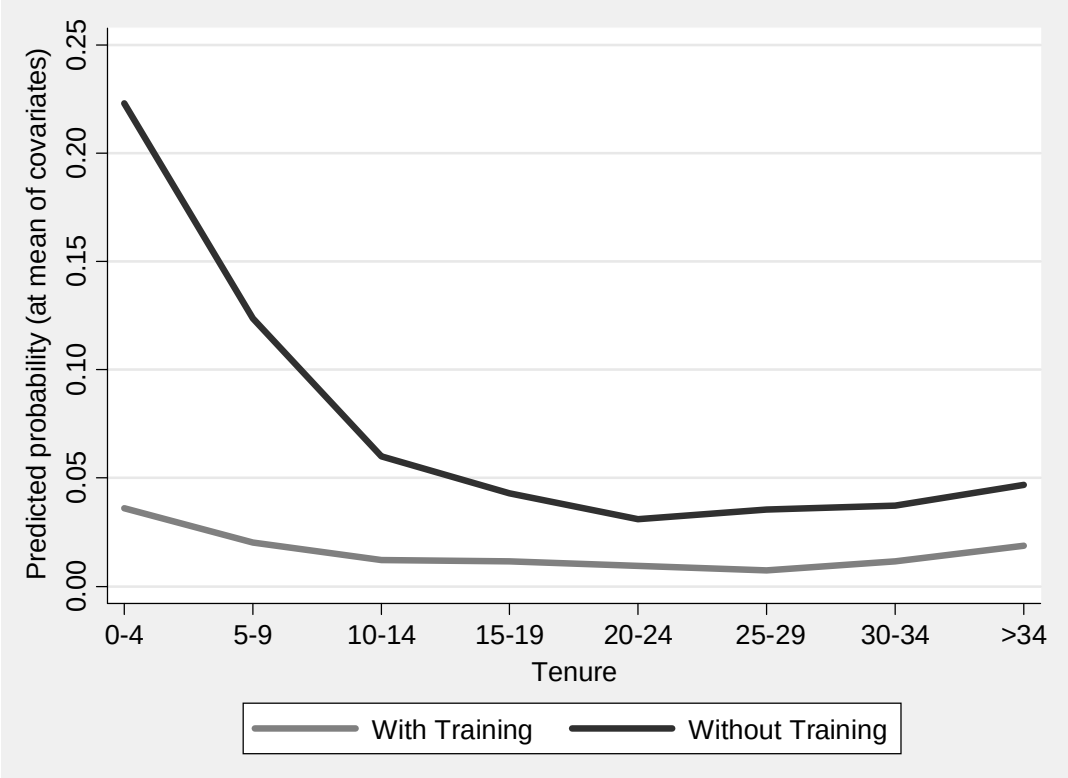

Figure 5.2: Predicted turnover probability for different tenure categories in dependence of training participation

the predicted turnover probabilities of employees without training proceed

above the probabilities for trained employees for all tenure groups. The difference is significant on the 1% level in all of the eight probit regressions.[16] However, the reduction in predicted turnover probability is much higher for lower-tenured employees. In a regression with interactions of training participation and tenure dummies we found a significant higher reduction in turnover probability due to training for employees with 0-4 years of tenure (on the 1% level) and for employees with 5-9 years of tenure (on the 5% level) compared to the middle-tenured group of 15-19 years of tenure. The finding supports our hypothesis that training indeed has an effect on the turnover probability of employees, especially for those which are new to the firm.

5.5.4 Robustness Tests for Predicting Turnover Probability

In a robustness check for the impact of training on turnover, we introduce further proxy variables for productivity to counter the argument that the salary percentile might be a weak proxy variable for productivity. We use two additional proxy variables: First, we observe annual bonus payments for managerial employees which are tied to individual performance evaluations in this firm. Second, for non-managerial employees, we observe overtime hours as described above. In two separate probit regressions, one for each employee group (managerial and non-managerial), we additionally control for these productivity measures and for the percentile position of the respective variable in the peer group. Table 5.10 shows that the effect of training participation remains robust in both specifications. There is a negative correlation of training in periods t and $t-1$ with the turnover rate in $t+1$ for both employee groups.

Like above, in the second robustness check separate probit regressions for the low and highly productive employees are applied to rule out that the effect is driven by one group**.** As productivity measures we use the three introduced proxies (salary, bonus and overtime percentile) and divide the subsample into a group above and below the median. The results are reported

[16] The regressions are available upon request.

Dependent variable:	Turnover$_{t+1}$			
	Managerial		Non-managerial	
	(1)	(2)	(3)	(2)
Training$_t$	-0.039***	-0.017***	-0.046***	-0.028***
	(0.004)	(0.005)	(0.002)	(0.004)
Training$_{t-1}$		-0.029***		-0.023***
		(0.006)		(0.004)
Percentile Bonus$_t$	-0.014	0.010		
	(0.009)	(0.015)		
Ln(Bonus)$_t$	0.021**	-0.052**		
	(0.009)	(0.022)		
Percentile Overtime$_t$			-0.020***	-0.011
			(0.006)	(0.008)
Overtime hours$_t$			0.000	0.000
			(0.000)	(0.000)
Observations	7,701	3,131	18,792	7,504
Pseudo R^2	0.099	0.107	0.243	0.292
Log likelihood	-1048.5051	-404.63875	-3444.1622	-1162.819

Marginal effects reported. Additional controls: see table 5.9

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

Table 5.10: Robustness Checks for the Effect of Training Participation on Turnover

in table 5.11 and show that the coefficients of training participation in t or $t-1$ are negative for all subsamples. In the regressions with the separation based on the median salary and median bonus payment compared to the peer group, we see that the coefficient is higher for the below-median group. But only the difference between the above and below-median group in the salary percentile regressions is significant on the 10% level (as tested in a separate regression with interaction terms). Overall, the result is robust for all subsamples.

Dep. var.:	Turnover$_{t+1}$					
	Salary percentile (all employees)		Bonus percentile (managerial)		Overtime percentile (non-managerial)	
	$>$ median	$\leq$ median	$>$ median	$\leq$ median	$>$ median	$\leq$ median
Training$_t$	-0.015***	-0.047***	-0.013**	-0.027***	-0.024***	-0.026***
	(0.003)	(0.007)	(0.006)	(0.009)	(0.004)	(0.004)
Training$_{t-1}$	-0.023***	-0.039***	-0.020***	-0.047***	-0.013***	-0.025***
	(0.003)	(0.007)	(0.006)	(0.012)	(0.004)	(0.005)
Obs.	7,261	3,798	1,965	1,129	2,629	5,186
Pseudo R^2	0.099	0.332	0.091	0.154	0.311	0.296

Marginal effects reported. Additional controls: see table 5.9
In column (3) to (6) additional controls for salary percentile. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 5.11: Robustness Checks for the Effect of Training Participation on Turnover

We also applied a kernel density matching approach. The propensity score is computed based on a probit regression predicting training participation (table 5.15 in the appendix). We used the same observable covariates as above except overtime and absence variables which are only available for non-managerial employees. Again, to capture possible selection of the same employees into training, we included the dummy for participation in $t-1$ in the probit model. Hence, we can only apply a matching approach for training participation in t and turnover probability in $t+1$ as we only have three years of data. The last column of table 5.14 (in the appendix) shows that after the matching procedure there are no significant differences in variables between the matched control group and the treated group. The pseudo R^2

of the probit regressions is reduced to 0 and the overlap condition is fulfilled (see figure 5.5 in the appendix). Hence, matching was successful and allows us to compare the outcome variable 'turnover$_{t-1}$' between the treated and untreated group. Table 5.12 shows that trained employees are about 4% less likely to leave the firm compared to their matched control group which supports the result from above.

Outcome variable: turnover$_{t+1}$	Obs	Difference	Bootstrap S.E.	P>|z|
ATT of Training$_t$	9964	-.0416***	0.006	0.000

Epanechnikov Kernel matching with bandwidth 0.06. Standard errors are based on bootstrapping with 200 replications. *** p<0.01, ** p<0.05, * p<0.1.

Table 5.12: Propensity Score Matching Result - Training in t and t-1 on Turnover Probability

5.6 Conclusion

Analyzing the effects of intra-firm training on employee's absence behavior and turnover probability based on personnel records of a multinational company, we find little support for human capital theoretic predictions. We rather observe that employees respond to training most probably due to reciprocal motives. Applying fixed effects regressions, we can show that trained employees are less absent in the year of training. But there is no persistent effect of reduced absence in the year after training. Furthermore, we find a negative effect of training participation on turnover in the subsequent year and two years after training. This suggests that besides strengthening the skills of the workforce, training increases the loyalty of employees and thereby even helps to retain qualified employees. In addition, the effect is stronger for employees with lower levels of firm tenure indicating that the training investment is especially helpful for firms to retain newly hired employees.

To further strengthen the positive effect on employees' perception of the firm when being trained**,** data of the firm's employee survey from 2007 was analyzed. Employees were asked if their unit "offers them good continuing education opportunities". We find a positive and significant correlation of

0.357 (p=0.0352) between average training participation in the work unit in $t-1$ and the outcome of this item based on 35 unit observations. This result is consistent with our interpretation and shows that employees express their belief in good continuing training opportunities of the firm when being trained more.

The results from our analysis may serve as a further explanation why firms invest in general training at all. Besides labor market frictions as addressed by Acemoglu and Pischke (1999), anticipating that employees positively react to training may lead to larger general training investments of firms in practice.

Our data set is unique because it guarantees a homogenous understanding of training policies and also includes reliable information from personnel records on the effort indicator sickness absence and on turnover. We are furthermore able to use the relative position of employees in the income and overtime distribution to control for time-varying selection effects into training and also apply matching procedures.

However, our study has some limitations. First, further research should try to collect more years of company records in order to better evaluate potential long-term effects of training. Especially for the analysis of turnover, statistical methods such as hazard rate models could then be applied. Second, our results may only partly be generalizable because we use a single firm data set. Hence, more company data sets should be collected in the future to test the reliability of the results. It might especially be helpful to match survey answers on employee level to personnel records in order to get a deeper understanding of the mechanism behind the individual reactions to firm-sponsored training.

5.7 Appendix of Chapter 5

Dependent variable:	Training participation$_t$	Training participation$_{t-1}$
	(1)	(2)
Training participation$^{a)}_{t-1/t-2}$	0.878***	0.814***
	(0.026)	(0.038)
Absence hours$_{t-1/t-2}$	-0.000	-0.000*
	(0.000)	(0.000)
Female	0.104***	0.135***
	(0.032)	(0.046)
Age$_{t/t-1}$	-0.013***	-0.016***
	(0.003)	(0.003)
Tenure$_{t/t-1}$	-0.007***	-0.004
	(0.002)	(0.003)
Salary percentile$_{t/t-1}$	-0.172***	-0.288***
	(0.0.065)	(0.097)
Ln(Salary)$_{t/t-1}$	0.254	-0.010
	(0.155)	(0.263)
Overtime percentile$_{t/t-1}$	0.104	-0.017
	(0.065)	(0.097)
Overtime hours$_{t/t-1}$	0.002***	0.003***
	(0.000)	(0.001)
Number of trainings supervisor$_{t/t-1}$	0.208***	0.204***
	(0.011)	(0.015)
Observations	14350	6529
Log likelihood	-6687.9734	-3235.1244
Pseudo R^2	0.2063	0.2014

$^{a)}$The first part of the subscript refers to the utilized variable in column (1) and the second part to the variable in column (2). Further controls are business unit, year and salary level dummies. No marginal effects reported.. *** p<0.01, ** p<0.05, * p<0.1.

Table 5.13: Probit Regression - Estimating Propensity Scores of Participating in Training in t and t-1 for Absent Hours

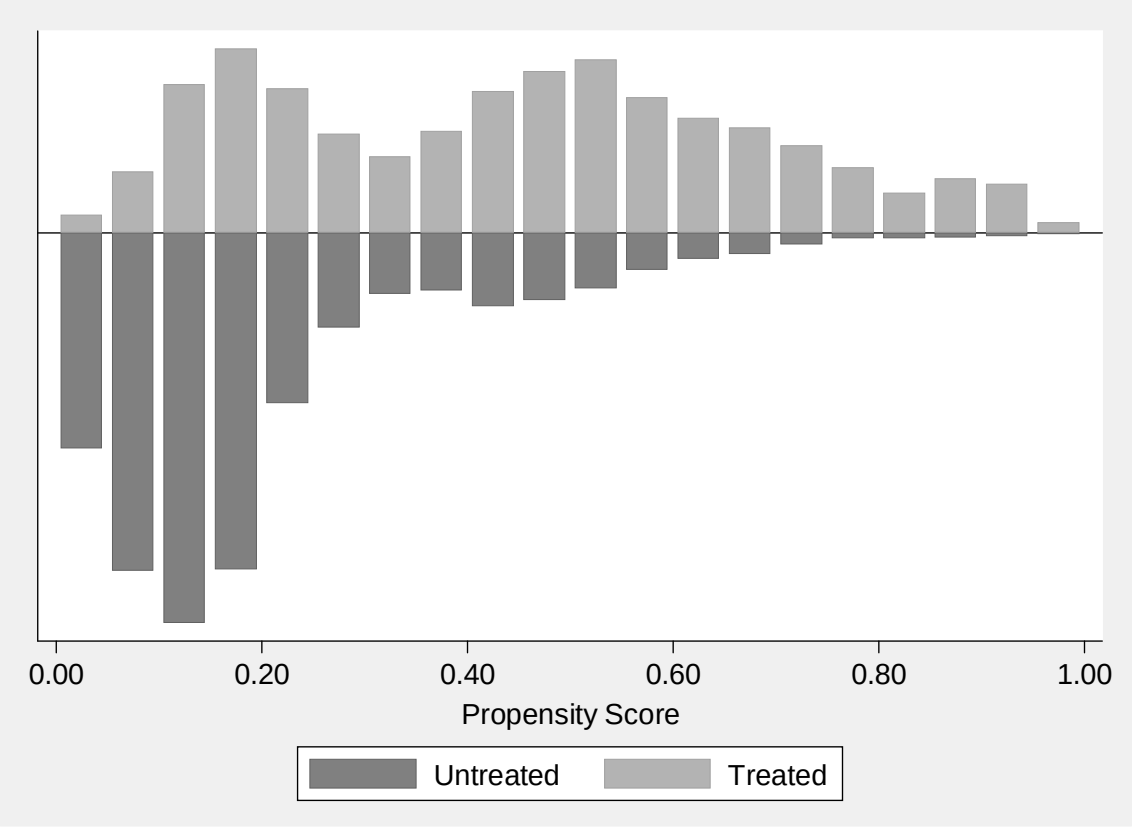

Figure 5.3: Propensity Scores Distribution for Matching Estimation of Training in t on Absence Hours in t

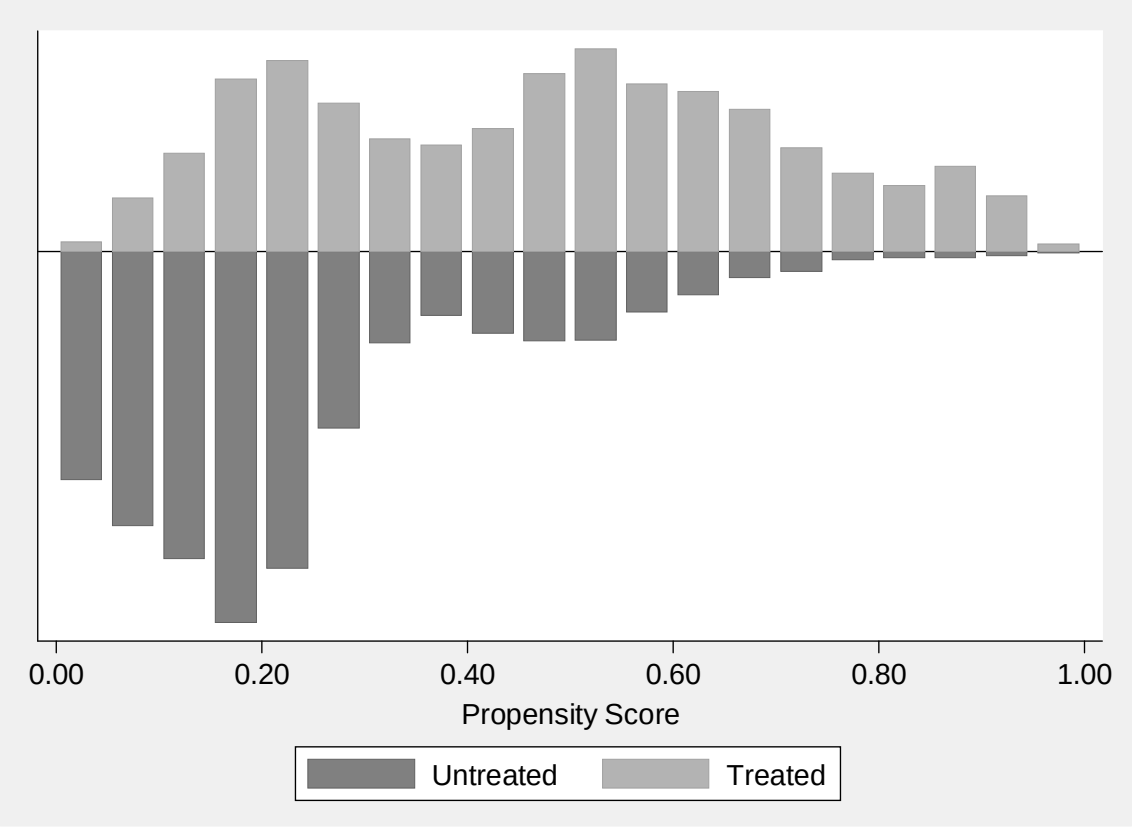

Figure 5.4: Propensity Score Distribution for Matching Estimation of Training in $t-1$ on Absence in t

	Absent hours$_t$				Turnover$_{t+1}$	
	Training in t		Training in $t-1$		Training in t	
	(1)		(2)		(3)	
	Before Matching	After Matching	Before Matching	After Matching	Before Matching	After Matching
T-test of equal means[a)]						
1%-level	24	0	18	0	15	0
5%-level	25	0	20	0	16	0
10%-level	25	1	21	1	16	0
Number of variables with standardized bias						
<1%	0	11	0	4	2	8
1% until <3%	1	8	0	11	1	9
3% until <5%	0	6	1	5	1	2
5% until <10%	5	0	3	2	0	0
≧10%	19	0	18	0	15	0
Pseudo R^2	0.206	0.002	0.201	0.004	0.191	0.001

[a)] Shown is the number of variables that differ significantly between treated and controls. Decision based on a simple t-test of means. There are 25 covariates used in the estimations of (1), 22 in (2) and 19 in (3).

Table 5.14: Matching Quality

Dependent variable:	Training participation$_t$
Training participation$_{t-1}$	0.817***
	(0.030)
Female	0.083**
	(0.038)
Age$_t$	-0.020***
	(0.003)
Tenure$_t$	-0.002
	(0.003)
Salary percentile$_t$	-0.135**
	(0.064)
Ln(Salary)$_t$	-0.400***
	(0.102)
Number of trainings supervisor$_t$	0.185***
	(0.012)
Observations	9964
Log likelihood	-5154.1372
Pseudo R^2	0.1913

Further controls are business unit, year and salary level dummies. No marginal effects reported.
*** p<0.01, ** p<0.05, * p<0.1.

Table 5.15: Probit Regression - Estimating Propensity Scores of Participating in Training in t and t-1 for Turnover Probability

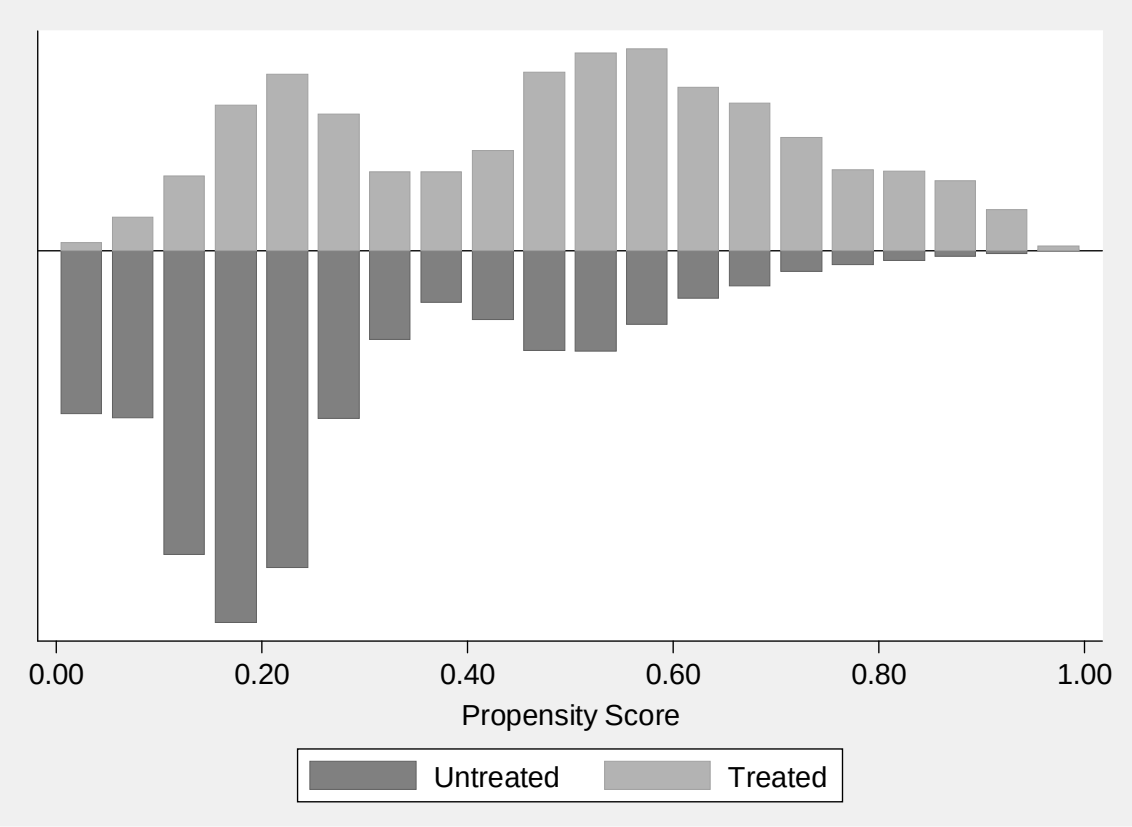

Figure 5.5: Propensity Score Distribution for Matching Estimation of Training in t on Turnover in $t+1$

Bibliography

Abbink, K., B. Irlenbusch, and E. Renner (2006). Group size and social ties in microfinance institutions. *Economic Inquiry 44*, 614–628.

Acemoglu, D. and J. Pischke (1999). Beyond becker: Training in imperfect labor markets. *Economic Journal 109*, 112–142.

Adams, R. B. and D. Ferreira (2009). Women in the boardroom and their impact on governance and performance. *Journal of Financial Economics 94*, 291–309.

Addison, J. T. and C. R. Belfield (2008). The determinants of performance appraisal systems: A note (do brown and heywoodt's results for australia hold up for britain?). *British Journal of Industrial Relations 46*, 521–531.

Akerlof, G. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics 97*, 543–69.

Anderson, S. W., H. C. Dekker, and K. L. Sedatole (2010). An empirical examination of goals and performance-to-goal following the introduction of an incentive bonus plan with participative goal setting. *Management Science 56(1)*, 90–109.

Baiman, S. and M. Rajan (1995). The informational advantage of discretionary bonus schemes. *The Accounting Review 70*, 557–579.

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

Barmby, T. A., C. D. Orme, and J. G. Treble (1991). Worker absenteeism: An analysis using microdata. *The Economic Journal 101*, 214–229.

Barrett, A. and J. O'Connell (2001). Does training generally work? the returns to in-company training. *Industrial and Labor Relations Review 54*, 647–662.

Barron, J., M. Berger, and D. Black (1997). How well do we measure training? *Journal of Labor Economics 15*, 507–528.

Barron, J., D. Black, and M. Loewenstein (1989). Job matching and on-the-job training. *Journal of Labour Economics 7*, 1–19.

Barron, J., D. Black, and M. Loewenstein (1993). Gender differences in training, capital and wages. *Journal of Human Resources 28(2)*, 343–364.

Barron, J. M., M. C. Berger, and D. A. Black (1998). Do workers pay for on-the-job training? *The Journal of Human Resources 34(2)*, 235–252.

Bartel, A. (1994). Productivity gains from the implementation of employee training programs. *Industrial Relations 33*, 411–425.

Bartel, A. (2000). Measuring the employert's return on investments in training: Evidence from the literature. *Industrial Relations 39*, 502–524.

Bartel, A. P. (1995). Training, wage growth, and job performance: Evidence from a company database. *Journal of Labour Economics 13*, 401–425.

Batt, R. (2002). Managing customer services: Human resource practices, quit rates, and sales growth. *Academy of Management Journal 45(3)*, 587–597.

Baumeister, R. F. and M. R. Leary (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin 117*(3), 497–529.

Becker, G. S. (1957). *The Economics of Discrimination*. University of Chicago Press.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of Political Economy 70*, 9–49.

Benedict, M. and E. Levine (1988). Delay and distortion: Tacit influences on performance appraisal effectiveness. *Journal of Applied Psychology 73(3)*, 507–514.

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

Berger, J., C. Harbring, and D. Sliwka (2010). Performance appraisals and the impact of forced distribution: An experimental investigation. *IZA Discussion Paper No. 5020*, 1–44.

Bernadin, H., D. Cooke, and P. Villanova (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology 85(2)*, 232–234.

Bjerk, K. (2008). Glass ceiling or sticky floors? statistical discrimination in a dynamic model of hiring and promotion. *The Economic Journal 118*, 961–982.

Black, S. and L. Lynch (1996). Human capital investments and productivity. *American Economic Review 86*, 263–267.

Bol, Jasmijn, C., T. M. Keune, E. M. Matsumura, and J. Y. Shin (2010). Supervisor discretion in target setting: An empirical investigation. *The Accounting Review 85(6)*, 1861–1886.

Bol, J. C. (2011). The determinants and performance effects of managersŠ performance evaluation biases. *The Accounting Review 86(5)*, 1549–1575.

Booth, A. L., M. Francesconi, and J. Frank (2003). A sticky floors model of promotion, pay, and gender. *European Economic Review 47*, 295–322.

Brandts, J. and C. Solà (2010). Personal relations and their effect on behaviour in an organizational setting: An experimental study. *Journal of Economic Behaviour and Organization 73(2)*, 246–253.

Bretz, R., G. Milkovich, and W. Read (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management 18*, 312–352.

Breuer, K. (2011). Are women better leaders? - an empirical investigation of gender differences in manager's evaluation behavior. *mimeo.*, 1–29.

Breuer, K. and P. Kampkötter (2011a). Do employees reciprocate to intra-firm training? an analysis of absenteeism and turnover probability. *mimeo.*, 1–37.

Breuer, K. and P. Kampkötter (2011b). The effects of intra-firm training on earnings and job performance - evidence from a large german company. *mimeo*, 1–25.

Breuer, K., P. Nieken, and D. Sliwka (2010). Social ties and subjective performance evaluations - an empirical investigation. *IZA Discussion Paper No. 4913*, 1–26.

Breuer, K. and J. Zimmermann (2011). Determinants and effects of target agreement systems: An empirical investigation of german firms. *mimeo.*, 1–31.

Brown, M. and J. Heywood (2005). Performance appraisal systems: Determinants and change. *British Journal of Industrial Relations 43*, 659–679.

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(2)*, 387–502.

Caliendo, M. and S. Kopeinig (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys 22(1)*, 31–72.

Cardy, R. L. and G. H. Dobbins (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. *Journal of Applied Psychology 71*(4), 672 – 678.

Carter, D. A. (2003). Corporate governance, board diversity, and firm value. *The Financial Review 38*, 33–53.

Catalyst (2010). Fortune 500 women board directors. *Research Reports: Catalyst Census*, 1–2.

Cedefop (2010). Employer-provided vocational training in europe. evaluation and interpretation of the third continuing vocational training survey. *European Centre for the Development of Vocational Training Research Paper No. 2*.

Conrads, J., B. Irlenbusch, R. M. Rilke, and G. Walkowitz (2011). Lying and team incentives. *IZA Discussion Paper No. 5968*, 1–10.

Costa, P. and R. McCrae (1992). *Revided NEO Personality Inventory (NEO PI-T) and NEO Five Factor Inventory. Professional Manual.* Odessa, Florida: Psychological Assessment Resources.

Croson, R. and U. Gneezy (2009). Gender differences in preferences. *Journal of Economic Literature 47:2*, 448–474.

Davies-Netzley, S. (1998). Women above the glass ceiling: Perceptions on corporate mobility and strategies for success. *Gender and Society 12*, 339–355.

Dearden, L., H. Reed, and J. van Reenen (2006). The impact of training on productivity and wages: Evidence from british panel data. *Oxford Bulletin of economics and statistics 68*, 397–421.

Dohmen, T., A. Falk, D. Huffmann, U. Sunde, J. Schupp, and W. G.G. (2006). Individual risk attitudes: New evidence from a large, representative, experimentally-validated survey. *CEPR Discussion Paper No. 5517*, 1–56.

Dreber, A. and M. Johannesson (2008). Gender differences in deception. *Economics letters 99(1)*, 197–199.

Drucker, P. F. (1954). *The Practice of Management*. Harper, New York.

Eagly, A., M. Johannesen-Schmidt, and M. van Engen (2003). Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing women and men. *Psychological Bulletin 129*, 569–591.

Eckel, C. and P. Grossman (1996). The relative price of fairness: Gender differences in a punsihment game. *Journal of Economic Behavior & Organization 30(2)*, 143–158.

Engellandt, A. and R. T. Riphahn (2011). Evidence on incentive effects of subjective performance evaluations. *Industrial and Labor Relations Review 64(2)*, 241–257.

Erhardt, N. L., J. D. Werbel, and C. B. Shrader (2003). Board of director diversity and firm financial performance. *Corporate Governance 11(2)*, 102–111.

Fahr, R., C. Schäfer, and D. Sliwka (2010). The performance effects of a training program - an econometric case study. *University of Paderborn Working Paper*.

Fairris, D. (2004). Internal labor markets and worker quits. *Industrial Relations 43(3)*, 573–594.

Falk, A. and M. Knell (2003). Choosing the joneses: Endogenous goals and reference standards. *The Scandinavian Journal of Economics 3*, 417–435.

Fehr, E. and S. Gächter (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives 14*, 159–181.

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

Fischer, G., F. Janik, D. Müller, and A. Schmucker (2008). The iab establishment panel - from sample to survey to projection. *FDZ Methodenreport 01/2008*, 1–38.

Fitzenberger, B. and G. Muehler (2011). Dips and floors in workplace training: Using personnel records to estimate gender differences. *ZEW Discussion Paper 110-023*, 1–37.

Freeman, R. B. and E. P. Lazear (1995). *Works Councils: Consultation, Representation, and Cooperation in Industrial Relations*, Chapter An Economic Analysis of Works Councils, pp. 27–52. University of Chicago Press.

Friebel, G. and P. Seabright (2011). Do women have longer conversations? telephone evidence of gendered communication strategies. *Journal of Economic Psychology 32*, 348–356.

Furnham, A. and P. Stringfield (2001). Gender differences in rating reports: female managers are harsher raters, particularly of males. *Journal of Managerial Psychology 16(4)*, 281–288.

García, M. U. (2005). Training and business performance: The spanish case. *International Journal of Human Resource Management 16(9)*, 1691–1710.

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.

Glaeser, E. L., D. I. Laibson, J. A. Scheinkman, and C. L. Soutter (2000). Measuring trust. *The Quarterly Journal of Economics 115*, 811–846.

Gneezy, U. (2004). Do high wages lead to high profits? an experimental study of reciprocity using real effort. *Working Paper, University of Chicago, Graduate Business School*, 1–31.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review 95*, 384–394.

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

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology 6*, 1360–1380.

Grote, R. C. (2005). *Forced Ranking*. Harvard Business School Press.

Grund, C. and D. Sliwka (2009). The anatomy of performance appraisals in germany. *The International Journal of Human Resource Management 20(10)*, 2049–2065.

Haines, V. Y., P. Jalette, and K. Larose (2010). The influence of human resource management practices on employee voluntary turnover rates in the canadian non governmental sector. *Industrial and Labor Relations Review 63(2)*, 228–246.

Hale, R. and P. Whitlam (1998). *Target setting and goal achievement: A practical guide for managers*. Kogan Page Limited, London.

Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management 20(4)*, 737–756.

Heath, C., R. P. Larrick, and W. G. (1999). Goals as reference points. *Cognitive Psychology 38*, 79–109.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998). Characterizing selection bias using experimental data. *Econometrica 66(5)*, 1017–1098.

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics 10*, 74–91.

Holt, C. and S. Laury (2002). Risk aversion and incentive effects. *American Economic Review 92(5)*, 1644–1655.

Howard, J. L. and D. D. Frink (1996). The effects of organizational restructure on employee satisfaction. *Group & Organization Management 21(3)*, 278–303.

Hsiao, C. (2003). *Analysis of Panel Data*. Cambridge University Press.

Ichino, A. and R. Riphahn (2005). The effect of employment protection on worker effort: Absenteeism during and after probation. *Journal of the European Economic Association 1*, 120–143.

Ichniowski, C., K. Shaw, and G. Prennushi (1997). The effects of human resource management pratices on productivity. *American Economic Review 87*, 291–313.

Jawahar, I. and C. Williams (1997). Where all the children are above average: the performance appraisal purpose effect. *Personnel Psychology 50*, 905–925.

John, O. and S. Srivastava (1999). *The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives.* in Handbock of Personality Psychology.

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

Jones, M. K., R. J. Jones, P. L. Latreille, and P. J. Sloane (2009). Training, job satisfaction, and workplace performance in britain: Evidence from wers 2004. *Labour 23 (Special Issue)*, 139–175.

Jovanovic, B. (1979). Job matching and the theory of turnover. *Journal of Political Economy 87*, 972–990.

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

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

Kaur, S., M. Kremer, and S. Mullainathan (2010). Self-control at work: Evidence from a field experiment. *mimeo*, 1–76.

Kingstorm, P. O. and L. E. Mainstone (1985). An investigation of the rater-ratee acquaitance and rater bias. *Academy of Management Journal 28*(3), 641 – 653.

Koch, A. and J. Nafziger (2009). Motivational goal bracketing. *IZA Discussion Paper No. 4471*, 1–30.

Koch, A. and J. Nafziger (2011). Self-regulation through goal setting. *Scandinavian Journal of Economics 113(1)*, 212–227.

Konrad, A. M., J. E. Ritchie, P. Lieb, and E. Corrigall (2000). Sex differences and similarities in job attribute preferences: A meta-analysis. *Psychological Bulletin 126*, 593–641.

Krueger, A. and C. Rouse (1998). The effect of workplace education on earnings, turnover and job performance. *Journal of Labor Economics 16*, 61–94.

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

Latham, G. P. and E. A. Locke (1990). *A Theory of Goal Setting and Task Performance*. Prentice-Hall: Engelwood Cliffs, New Jersey.

Lazear, E. (1979). Why is there mandatory retirement? *Journal of Political Economy 87*, 1261–1284.

Lazear, E. and S. Rosen (1990). Male-female wage differentials in job ladders. *Journal of Labor Economics 8(1)*, S106–S123.

Lazear, E. P. (1981). Agency, earnings profiles, productivity, and hours restrictions. *American Economic Review 71(4)*, 606–620.

Lazear, E. P. and M. Gibbs (2009). *Personnel Economics in Practice*. John Wiley & Sons.

Lechner, M. (1999). Earnings and employment effects of continous off-the-job training in east germany after unification. *Journal of Business Economic Studies 17(1)*, 74–90.

Lee, C. H. and N. T. Bruvold (2003). Creating value for employees: investment in employee development. *International Journal of Human Resource Management 14(6)*, 981–1000.

Lefkowitz, J. (2000). The role of interpersonal affective regard in supervisory performance ratings: A literature review and proposed causal model. *Journal of Occupational & Organizational Psychology 73*(1), 67 – 85.

Levine, D. I. (1993). Worth waiting for? delayed compensation, training, and turnover in the unisted states and japan. *Journal of Labor Economics 11(4)*, 724–752.

Levy, P. E. and J. Williams (2004). The social context of performance appraisal: a review and framework for the future. *Journal of Management 30*, 881–906.

Lincoln, J. R. and A. L. Kalleberg (1996). Commitment, quits, and work organization in japanese and u.s. plants. *Industrial and Labor Relations 50(1)*, 39–59.

Locke, E. and G. Latham (2002). Building a practically useful theory of goal setting and task motivation. *American Psychologist 57*, 705–717.

Locke, E. A. (1968). Toward a theory of task motivation and incentives. *Organizational Behavior and Human Performance 3*, 157–189.

Lynch, L. (1992). Private sector training and the earnings of young workers. *American Economic Review 82(1)*, 299–312.

MacLeod, W. (2003). Optimal contracting with subjective performance evaluation. *American Economic Review 93, 1*, 216–240.

Medoff, J. L. and K. G. Abraham (1980). Experience, performance, and earnings. *The Quarterly Journal of Economics 95*, 703–36.

Milkovich, G. T. and A. K. Wigdor (1991). *Pay for Performance*. National Academy Press.

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

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

Murphy, K. J. (2001). Performance standards in incentive contracts. *Journal of Accounting and Economics 30*, 245–278.

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

Napier, N. and G. Latham (1986). Outcome expectancies of people who conduct performance appraisals. *Personnel Psychology 39*, 827–837.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *Quarterly Journal of Economics 122(3)*, 1067–1101.

Parent, D. (1999). Wages and mobility: The impact of employer-provided training. *Journal of Labor Economics 17*, 298–317.

Picchio, M. and J. van Ours (2011). Retaining through training: Even for older workers. *IZA Discussion Paper No. 5591*, 1–27.

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

Prendergast, C. J. (2002). Uncertainty and incentives. *Journal of Labor Economics 20*, 115–37.

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

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

Rablen, M. D. (2010). Performance targets, effort and risk-taking. *Journal of Economic Psychology 31*, 687–697.

Riphahn, R. (2004). Employment protection and effort among german employees. *Economics Letters 85*, 353–357.

Rodgers, R. and J. Hunter (1991). Impact of management by objectives on organizational productivity. *Journal of Applied Psychology Monograph 76*, 322–336.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika Trust 70(1)*, 41–55.

Roth, P., K. Purvis, and P. Bobko (2011). A meta-analysis of gender group differences for measures of job performance in filed studies. *Journal of Management 37(2)*, 1–21.

Schmitt, D., A. Realo, M. Voracek, and J. Allik (2008). Why can't man be more like a women? sex differences in big five personality traits across cultures. *Journal of Personality and Social Psychology 94*, 168–182.

Shaw, J. D., J. E. Delery, G. D. Jenkins Jr., and N. Gupta (1998). An organizational-level analysis of voluntary and involuntary turnover. *Academy of Management Journal 41(5)*, 511–525.

Shore, T. and A. Tashchian (2002). Accountability forces in performance appraisal: Effects of self-appraisal information, normative information, and task performance. *Journal of Business and Psychology 17(2)*, 261–274.

Smith, J. and P. Todd (2005). Does matching overcome lalondet's critique of non-experimental estimators? *Journal of Econometrics 125(1)*, 305–353.

Terpstra, D. E. and E. L. Rozell (1994). The relationship of goal setting to organizational profitability. *Group & Organization Management 19*, 285–294.

Van Praag, B. and A. Ferrer-i Carbonel (2004). *Happiness Quantified: A Satisfaction Calculus Approach*. Oxford, Oxford University Press.

Varma, A., A. S. Denisi, and L. H. Peters (1996). Interpersonal affect and performance appraisal: A field study. *Personnel Psychology 49*(2), 341 – 360.

Varma, A. and L. Stroh (2001). The impact of same-sex lmx dyads on performance evaluations. *Human Resource Management 40*, 309–320.

Veum, J. (1995). Sources of training and their impact on wages. *Industrial and Labor Relations 48(4)*, 812–826.

Welch, J. (2003). *Jack- Straight from the Gut*. Warner Books.

Winkelmann, R. (1999). Wages, firm size and absenteeism. *Applied Econonomic Letters 6*, 337–341.

Yariv, E. (2006). Mum effects: principals' reluctance to submit negative feedback. *Journal of Managerial Psychology 21(6)*, 533–546.

Zwick, T. (2006). The impact of training intensity on establishment productivity. *Industrial Relations 45*, 26–46.