Breuer, Kathrin; Nieken, Petra; Sliwka, Dirk

Working Paper

Social ties and subjective performance evaluations: An empirical investigation

IZA Discussion Papers, No. 4913

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Breuer, Kathrin; Nieken, Petra; Sliwka, Dirk (2010) : Social ties and subjective performance evaluations: An empirical investigation, IZA Discussion Papers, No. 4913, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/36827

DISCUSSION PAPER SERIES

IZA DP No. 4913

Social Ties and Subjective Performance Evaluations: An Empirical Investigation

Kathrin Breuer
Petra Nieken
Dirk Sliwka

April 2010

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Social Ties and Subjective Performance Evaluations: An Empirical Investigation

Kathrin Breuer
University of Cologne

Petra Nieken
University of Bonn

Dirk Sliwka
University of Cologne and IZA

Discussion Paper No. 4913
April 2010

IZA

P.O. Box 7240
53072 Bonn
Germany

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

IZA Discussion Paper No. 4913
April 2010

ABSTRACT

Social Ties and Subjective Performance Evaluations: An Empirical Investigation

We empirically investigate possible distortions in subjective performance evaluations. A key hypothesis is that evaluations are more upward biased the closer the social ties between supervisor and appraised employee. We test this hypothesis with a company data set from a call center organization which contains not only subjective assessments but also several more objective measures of performance. Controlling for these performance measures, we find strong evidence that evaluations are upwards biased in smaller teams and some evidence that supervisors give better ratings to employees they themselves have evaluated before.

JEL Classification: D03, M52

Keywords: subjective performance evaluation, bias, social ties, team size, favoritism

Corresponding author:

Dirk Sliwka
University of Cologne
Herbert-Lewin-Str. 2
50931 Köln
Germany
E-mail: dirk.sliwka@uni-koeln.de

1 Introduction

In many jobs, not all aspects of employee performance are objectively measurable. Therefore, organizations frequently use subjective performance evaluations to measure the employees' contribution. Gibbs et al. (2003), for instance, have argued that the use of subjectivity in performance evaluation can strengthen incentive setting as more facets of the job can be appraised. On the other hand the use of subjective components in evaluations raises issues of rating bias which can cause substantial inefficiencies (see for instance Prendergast and Topel (1993), Murphy and Cleveland (1995), or Moers (2005)). 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 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 and thus to contribute to the progress of "*understanding how subjective assessments are made*" (Prendergast (1999)).

Our aim is to shed some light on the question whether and why subjective performance evaluations are distorted using a unique data set from a call center organization. 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. 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.

A key observation in the literature is that performance appraisals tend

to be too lenient. 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 tradeoff between accuracy and leniency and it directly results that evalutions are the more lenient, the stronger the supervisor's social preferences towards the evaluated subordinate. Based on this reasoning, we now argue that a closer social attachment between supervisor and subordinate should lead to better performance ratings even when there are no differences in true performance.

We use two proxies for social ties. First, we suppose that the strength of the personal relationship between supervisor and subordinate depends on the size of the group evaluated. We therefore analyze the effect of work unit size on the result of subjective evaluations and expect more lenient results for smaller units 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 of course important to stress that we control for objective measures of performance, employee specific fixed effects, as well as prior job experience to exclude that the results are driven simply by differences in productivity.

A key underlying assumption is of course that the frequency of interaction increases social attachment. There is quite substantial evidence backing this claim. In a very exhaustive psychological review on social attachment 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) for instance 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 (2006) study the effect of personal relations on distributive decisions and find discrimination against the subjects that are not personally known to the distributor.[1]

[1] 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)

The connection between the degree of acquaintance between rater and ratee or rating biases has also been discussed in the psychological literature (see for instance Cardy and Dobbins (1986), Varma et al. (1996), or Lefkowitz (2000)). Most studies are either laboratory experiments with students or they lack objective measures of performance. 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. They find a weak positive correlation between both and rating leniency in a cross section analysis.

In our study we use panel data on performance evaluations from a call center over 4 years. The investigated subjects are call-agents whose main task is to deal with service queries over the telephone from clients who bought technical products. We have information about the average handling time (AHT), so-called Transaction Monitoring (TM) scores and the days of absence. In the Transaction Monitoring process the quality of the agent's interaction with the client is assessed on the basis of a narrow defined requirement catalogue by an external monitor who is not the direct supervisor. Controlling for these performance measures, helps us to discover systematic distortions in the evaluation process. Moreover, as we have an (unbalanced) panel, the performance of a number of employees in the sample is evaluated by different supervisors at different points in time and also groups are rearranged frequently, we can control for unobserved heterogeneity in agents' and supervisor's characteristics.

Our results indeed show a significant negative influence of unit size on performance evaluations. In smaller groups where the personal contact between supervisor and employee is closer, the overall subjective assessment grades are significantly better. Furthermore we find that employees who have been assessed by the same supervisor before, on average receive better ratings than colleagues of the same tenure and who attained the same transaction monitoring scores.

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.

The remainder of the paper in organized as follows. Section 2 deals with the institutional background and section 3 with the empirical approach. We present the results in section 4. Section 5 concludes.

2 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 business activities of the company are organized in departments, of which we observe a total of 12 in the full sample over the years.[2] 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. It The evaluation data is stored in an internal database with the exact time period the

[2]We only look at the departments of the primary core business activity. Human Resources, Accounting, IT etc. are excluded.

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).

3 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 baseline specification:

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

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 attachment 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 and υ_t

year dummies. As the dependent variable is measured on an ordinal scale we additionally run ordered probit regressions.[3] To control for unobserved heterogeneity in personal characteristics of the employees we also estimate individual fixed and random effects models.

We apply two main indicators for social proximity in our analysis. First, group size is measured by the quantity of evaluations an assessor conducted per year. For every supervisor the absolute number of evaluations conducted per year is summed up in a variable called "Assessments per year". Secondly, a dummy variable is introduced indicating an appraisal being conducted by the same supervisor the year before as a proxy for the time of acquaintance.

Performance measures used as control variables are the average result of the Transaction Monitoring, the standardized 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.[4]

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

[3]Note that nearly 89% of the observations received a 3 ("fulfilled requirements") which affirms a "*managers' tendency to assign uniform ratings to employees*" (Murphy (1992)).

[4]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.

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 A1 in the appendix.

4 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 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 1: Distribution of appraisal grades by group size

Regression results regarding the effects of the number of assessed employees are shown in table 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 10%-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).

	OLS		Ordered Probit	
Overall appraisal	(1)	(2)	(3)	(4)
Assessments per year	-0.0034^{*}	-0.0051^{***}	-0.0170^{**}	-0.0262^{***}
	(0.0017)	(0.0015)	(0.0073)	(0.0064)
TM		0.0095^{***}		0.0468^{***}
		(0.0028)		(0.0091)
Days of absence		-0.0385^{***}		-0.1968^{***}
		(0.0107)		(0.0447)
Over 100% AHT		-0.0243		-0.1548
		(0.0268)		(0.1521)
Under 90% AHT		0.0104		0.0412
		(0.0366)		(0.1763)
New Assessor	-0.2200^{**}	-0.2631^{***}	-0.8878^{***}	-1.1415^{***}
	(0.0936)	(0.0802)	(0.3173)	(0.2561)
Female	-0.0710	-0.0662	-0.3352	-0.3512
	(0.0484)	(0.0479)	(0.2503)	(0.2874)
Tenure	0.0152^{***}	0.0188^{***}	0.0763^{***}	0.1163^{***}
	(0.0052)	(0.0059)	(0.0257)	(0.0282)
Constant	4.6491^{***}	3.5810^{***}		
	(0.5693)	(0.4330)		
Observations	520	520	520	520
R^2	0.096	0.158		
Pseudo Likelihood			-193.95661	-177.04632

Robust Standard errors in parentheses. Clustered on team level. Control variables include age,age^2, year dummies, the share of women and average team age.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 2: Number of Assessments: OLS and Ordered Probit

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 is insignificant. Having an assessor who has never rated before has also the

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 3 and confirm the previous observations in all specifications. The model predicts than a specific employee switching from a smaller to a larger group will receive an inferior evaluation even if his true performance is unaffected.

To evaluate the economic significance of the effects, we additionally conducted a probit analysis reporting the marginal effect of group size on the probability of receiving a good evaluation (i.e. receiving a either a 4 or 5). We include dummy variables indicating the particpation in a small (< 15) or a large group (≥ 30). As can be seen in table A2 in the appendix the probability of receiving a good grade is about 5.4% higher when being part of a small team in comparison to the reference group of a middle-sized team while there is no significant difference between large and middle sized teams.

To investigate our second hypothesis we now 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 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 bad grade 2 more than 5 times as often as supervisors who have evaluated the same employee before.

It is also interesting to compare changes in grades for given employees: 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

	Random effects		Fixed effects	
Overall appraisal	(1)	(2)	(3)	(4)
Assessments per year	-0.0037^{***}	-0.0053^{***}	-0.0074^{***}	-0.0079^{***}
	(0.0010)	(0.0011)	(0.0023)	(0.0024)
TM		0.0097^{***}		0.0077^{*}
		(0.0023)		(0.0044)
Days of absence		-0.0378^{**}		-0.0230
		(0.0151)		(0.0349)
Over 100% AHT		-0.0169		0.0714
		(0.0327)		(0.0654)
Under 90% AHT		0.0095		0.0014
		(0.0340)		(0.0714)
New Assessor	-0.2241^{***}	-0.2698^{***}	-0.3320^{**}	-0.3893^{**}
	(0.0697)	(0.0670)	(0.1560)	(0.1590)
Female	-0.0734^{**}	-0.0691^{*}		
	(0.0369)	(0.0362)		
Tenure	0.0150	0.0191^{**}	-0.0948^{*}	-0.1037^{**}
	(0.0095)	(0.0096)	(0.0544)	(0.0517)
Constant	4.6973^{***}	3.6272^{***}	8.4631^{**}	8.2989^{**}
	(0.4641)	(0.4771)	(3.5930)	(3.6324)
Observations	520	520	520	520
R^2	0.095		0.136	0.170

Robust Standard errors in parentheses. Control variables include age, age^2, year dummies, the share of women and average team age.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 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 4: Distribution of appraisal grades by "repeated assessment"

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%).

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 5. 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.

	OLS			Ordered Probit		
Overall appraisal	(1)	(2)	(3)	(4)	(5)	(6)
Repeated Appraisal Same Supervisor	0.0754*	0.0730*	0.0515	0.3643*	0.3664**	0.2836*
	(0.0404)	(0.0377)	(0.0332)	(0.1868)	(0.1746)	(0.1723)
Female	−0.0829	−0.0792	−0.0755	−0.3945	−0.4184	−0.4131
	(0.0530)	(0.0548)	(0.0541)	(0.2574)	(0.2914)	(0.3009)
Tenure	0.0180**	0.0238***	0.0226***	0.0855***	0.1268***	0.1246***
	(0.0076)	(0.0073)	(0.0064)	(0.0309)	(0.0288)	(0.0280)
TM		0.0064*	0.0069*		0.0331**	0.0353**
		(0.0036)	(0.0038)		(0.0138)	(0.0145)
Days of absence		−0.0410***	−0.0391***		−0.1860***	−0.1823***
		(0.0106)	(0.0104)		(0.0438)	(0.0454)
Over 100% AHT		−0.0284	−0.0298		−0.1456	−0.1646
		(0.0257)	(0.0253)		(0.1402)	(0.1414)
Under 90% AHT		0.0015	0.0013		0.0066	−0.0110
		(0.0338)	(0.0337)		(0.1641)	(0.1617)
New Assessor			−0.1702*			−0.6729**
			(0.0919)			(0.3110)
Constant	4.0407***	3.0328***	2.7758***			
	(0.7097)	(0.7854)	(0.6434)			
Observations	520	520	520	520	520	520
R^2	0.065	0.107	0.124			
Log Pseudolikelihood				−200.75498	−189.13028	−185.86686

Robust Standard errors in parentheses. Clustered on team level.

Control variables include age, ageš, year dummies, the share of women and average team age.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 5: 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 heterogeity. 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 measured 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 A1 and A2 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.

Hence, to control for unobserved heterogeneity and selection effects we therefore again ran random and fixed effects regressions (see table 6).

	Random Effects			Fixed Effects		
Overall appraisal	(1)	(2)	(3)	(4)	(5)	(6)
Repeated Appraisal Same Supervisor	0.0793**	0.0776**	0.0534	0.1361*	0.1815**	0.1596**
	(0.0375)	(0.0383)	(0.0389)	(0.0716)	(0.0719)	(0.0685)
Female	−0.0855**	−0.0825**	−0.0792**			
	(0.0388)	(0.0385)	(0.0382)			
Tenure	0.0177*	0.0238**	0.0230**	−0.1039	−0.1336**	−0.1351**
	(0.0096)	(0.0098)	(0.0097)	(0.0650)	(0.0607)	(0.0578)
TM		0.0065***	0.0070***		0.0092*	0.0101*
		(0.0020)	(0.0020)		(0.0053)	(0.0051)
Days of absence		−0.0399**	−0.0380**		−0.0069	−0.0051
		(0.0172)	(0.0163)		(0.0342)	(0.0340)
Over 100% AHT		−0.0204	−0.0226		0.0423	0.0439
		(0.0335)	(0.0333)		(0.0661)	(0.0648)
Under 90% AHT		0.0008	0.0003		−0.0088	−0.0111
		(0.0347)	(0.0344)		(0.0736)	(0.0724)
New Assessor			−0.1660**			−0.1027
			(0.0650)			(0.1243)
Constant	4.0198***	3.0325***	2.7805***	3.0612	4.9775	5.0499
	(0.4151)	(0.5004)	(0.4922)	(3.3238)	(3.4735)	(3.4994)
Observations	520	520	520	520	520	520
R^2				0.064	0.137	0.143
Number of persons	386	386	386	386	386	386

Robust Standard errors in parentheses.

Control variables include age, ageš, year dummies, the share of women and average team age.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 6: 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.

Finally, 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 A4, the standard deviation of assessments by the same supervisor is smaller rather than larger which also makes such a mechanism unplausible.

The results concerning both hypotheses 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 A3. But the effects of team size are more robust than those of repeated appraisals.

5 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 regression such that a given person with a given experience and performance measures receives lower ratings when moving to a larger team or when getting a new supervisor.

Our results indicate that firms must be cautious when using performance appraisal results to compare employees across departments. 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.

References

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

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.

Brandts, J. and C. Solà (2006). Personal relations and their effect on behaviour in an organizational setting: An experimental study. *Working Paper*, 1–29.

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.

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.

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.

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.

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

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.

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

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

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

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

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

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

6 Appendix

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	
Repeated Appraisal Same Supervisor (Dummy)	0.113	
Objective Performance Measures		
Result Transaction Monitoring (TM)	90.554	8.992
Over 100% AHT per group-year (Dummy)	0.462	
Under 90% of mean AHT per group-year (Dummy)	0.285	
Days of absence (standardized)	0.107	1.121
Individual Characteristics		
Tenure	2.754	1.988
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	
Dummy new assessor (1/0)	0.077	

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

Table A1: Descriptive Statistics

	Probit - Marginal Effects
Good grade (4 or 5) (Dummy)	
Small group (< 15) (Dummy)	0.0537^{**}
	(0.0448)
Large group (≥ 30) (Dummy)	-0.00592
	(0.0107)
TM	0.00189^{***}
	(0.000794)
Days of absence	-0.0178^{**}
	(0.00740)
Over 100% AHT (Dummy)	0.00702
	(0.0121)
Under 90% AHT (Dummy)	0.0162
	(0.0173)
New Assessor	-0.0134
	(0.00779)
Female	-0.00403
	(0.0103)
Observations	520
Pseudo Likelihood	-67.750413

Robust Standard errors in parentheses.

Control variables include age, age^2, year dummies, the share of women and average team age

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table A2: Probit Estimation - Marginal Probability of receiving a good grade (4 or 5)

	Random Effects		Fixed Effects	
Overall appraisal	(1)	(2)	(1)	(2)
Assessments per year	−0.00343***	−0.00520***	−0.00305*	−0.00692***
	(0.00102)	(0.00113)	(0.00172)	(0.00252)
Repeated Appraisal Same Supervisor (Dummy)	0.0735*	0.0336	0.179**	0.110
	(0.0392)	(0.0404)	(0.0722)	(0.0703)
Tenure	0.0213**	0.0185*	−0.122*	−0.111**
	(0.00979)	(0.00961)	(0.0626)	(0.0518)
TM	0.00801***	0.00961***	0.00751	0.00796*
	(0.00220)	(0.00232)	(0.00522)	(0.00445)
Days of absence	−0.0410**	−0.0383**	−0.0148	−0.0193
	(0.0167)	(0.0151)	(0.0344)	(0.0346)
Over 100% AHT (Dummy)	−0.0142	−0.0134	0.0558	0.0781
	(0.0333)	(0.0329)	(0.0669)	(0.0657)
Under 90% AHT (Dummy)	0.00792	0.0109	−0.00141	0.000701
	(0.0342)	(0.0339)	(0.0732)	(0.0707)
New Assessor		−0.263***		−0.314*
		(0.0686)		(0.165)
Average age per team	yes	yes	yes	yes
Female share per team	yes	yes	yes	yes
Age, Age^2, Yeardummies	yes	yes	yes	yes
Constant	3.699***	3.645***	6.100*	7.744**
	(0.486)	(0.480)	(3.447)	(3.543)
Observations	520	520	520	520
R^2	0.1243	0.1584	0.150	0.183
Number of persons	386	386	386	386

Robust Standard errors in parentheses.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table A3: Number of Assessments and Repeated Appraisal: Fixed Effects

Grades by new assessments	Mean	Sd
Different supervisor	2.9753	0.3558
Same supervisor	3.0333	0.2582

Table A4: Mean and Standard Deviation by "repeated assessment"

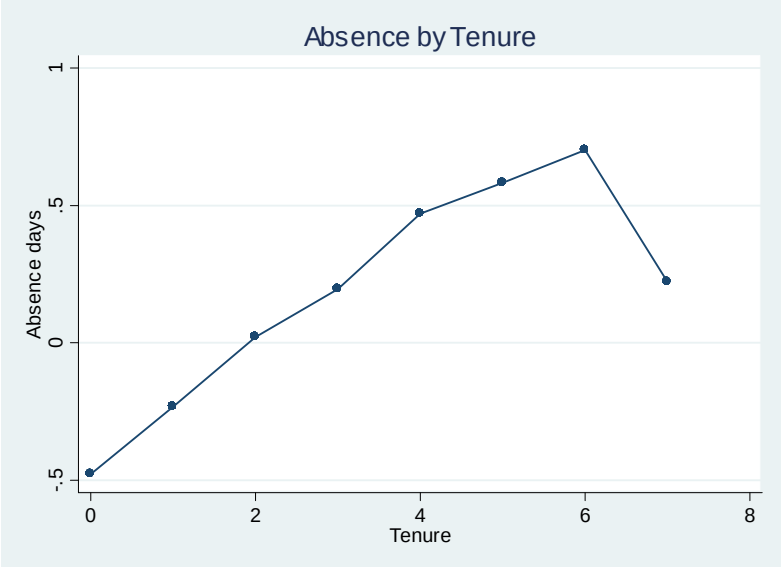

Figure A1: Absence in days (standardized) by years of tenure

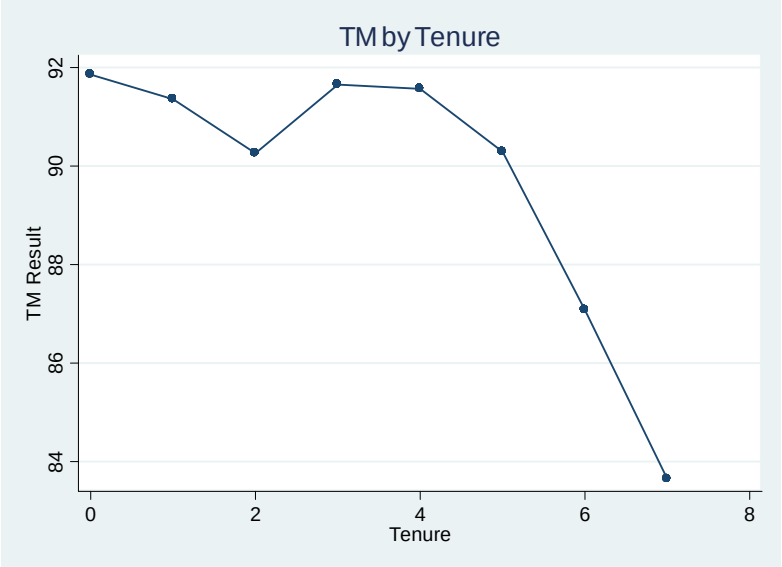

Figure A2: Transaction Monitoring by years of tenure