Social Ties and Subjective Performance Evaluations - An Empirical Investigation

Kathrin Breuer[*]
University of Cologne

Petra Nieken
University of Cologne

Dirk Sliwka
University of Cologne and IZA

December 2008

Abstract

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.

Key Words: Subjective Performance Evaluation, Bias, Social Ties, Team Size, Favoritism

[*] University of Cologne, Herbert-Lewin-Str. 2, 50931 Köln, Germany, tel: +49 221 470-3894, fax: +49 221 470-5078, e-mail: kathrin.breuer@uni-koeln.de, petra.nieken@uni-koeln.de, dirk.sliwka@uni-koeln.de.

1 Introduction

Many organizations use subjective performance evaluations to measure the employee's contribution to firm value because in many jobs employee performance is rarely objectively measurable (for an overview see Murphy and Cleveland (1995)). 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 bias which can cause substantial inefficiencies (compare Prendergast and Topel (1993) or Moers (2005)). In a subjective assessment "*human judges other humans*" (Milcovich and Wigdor (1991)) which for instance may open the door to favoritism, so that supervisors can exercise their personal preferences and bias the outcome of the evaluation. A biased performance evaluation can lead to an inefficient allocation of workers to certain tasks or jobs (Prendergast and Topel (1996)) and fail to identify training needs of employees when they are judged more leniently. Therefore there is a necessity of investigating potential distortions of subjective evaluations in a real organizational context, which we do here and thus contribute to the progress of "*understanding how subjective assessment are made*" (Prendergast (1999)).

We are able to shed some light into the question to what extent subjective performance evaluation is 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 attachment. A key feature of this data set is that besides a subjective evaluation we have a number of more objective measures of performance. Therefore, we are able to control for the main aspects of the true performance of an employee.

There are a series of consequences that can result from social attachment between supervisor and employee in the workplace.[1] Our key hypothesis here is that a closer social attachment between supervisor and employee leads to more lenient performance ratings. 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 assume a closer personal relationship to the supervisor for employees who have worked for the supervisor a longer period of time and therefore expect more lenient ratings for these. We argue that the extent of potential personal preferences and resulting biases depends on the intensity of the personal contact between supervisor and subordinate.

The connection between personal affect or the degree of acquaintance between rater and ratee and 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 lacking 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 and find a weak positive correlation between both and rating favorability controlling for objectively measurable sales productivity 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 major job is to answer queries from clients over the telephone. We have information about the average handling time (AHT), so-called Transaction Monitoring (TM) scores and the days of absence which we use as objective measures of performance. 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. By controlling for these "objective" performance measures, we can discover systematic distortions in the evaluation through supervisors. Moreover, as we have an (unbalanced)

[1] See for example also Bandiera et al. (2008) for a survey.

panel, the performance of some employees 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 show a significant negative influence of unit size on performance evaluations when controlling for objective performance measures. 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 receive better ratings than colleagues of the same tenure who attained the same objectively measurable performance.

2 Theoretical Background and Hypotheses

The key idea is illustrated with a simple version of Prendergast and Topel (1996) or Prendergast (2002)'s model of subjective performance evaluation.[2] A supervisor S has to evaluate the performance of an agent A_i. The supervisor observes the agent's actual performance s_i and gives a report r_i. The agent's utility depends on this evaluation and is given by $u_{A_i} = \alpha + \beta r_i$. We assume (i) that the supervisor cares to some extent for making an accurate report (ii) but her utility is also affected by the well-being of her subordinate. She therefore trades-off these two components of her utility when making the report, such that her utility is given by

$$u = \mu_i \cdot u_{A_i} - \nu \cdot (r_i - s_i)^2 .$$

Hence, μ_i measures how much the supervisor cares for the well-being of agent A_i or in other words is the degree of *social attachment* towards this agent. On the other hand, ν measures how much she cares for accuracy in her report. For instance, she is controlled with a certain probability and has

[2] We use this model only for illustration purposes. Theoretical work on the use of subjective evaluations in optimal contracts has for instance been done by Baker et al. (1994) or MacLeod (2003).

to pay a fine if her report turns out to be inaccurate or simply she has a guilty conscience if her report differs too much from the actual performance.

The supervisor maximizes her utility

$$\mu_i \cdot (\alpha + \beta r_i) - \nu \cdot (r_i - s_i)^2$$

yielding a report

$$r_i = s_i + \frac{\mu_i \beta}{2\nu}.$$

Hence, the performance report is biased upwards as long as $\mu_i > 0$ and the report is increasing in μ_i.

Of course, the personal preference of the supervisor is not measurable, but can be proxied by characteristics of the relationship between supervisor and employee. To do this we study the effect of the size of the work unit and of repeated assessments on rating elevation. A key assumption underlying the use of these indicators for social proximity is 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.[3]

To summarize, we hypothesize that there will be closer social ties in small teams for instance as supervisor and agent meet more frequently leading to higher values of μ_i:

Hypothesis 1: Employees in smaller units receive more lenient ratings.

[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"-treatments.

Moreover, when supervisor and agent know each other for a longer period of time we expect μ_i to be higher on average. As we cannot measure directly how long supervisor and agent know each other we use repeated assessments as a proxy, i.e. we argue that an employee who has been assessed by the same supervisor before has closer social ties to her than a colleague who is assessed by the supervisor for the first time. As we also should expect that employees with a longer tenure have build up more human capital we control for tenure as well as for the more objective performance measures:

Hypothesis 2**:** Employees who have been assessed by the same supervisor several times before receive more lenient ratings.

3 Empirical Investigation

3.1 Institutional Background

We investigate personnel data on 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.[4] 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, second level agents and first level agents. The communication coach is responsible to train the communication skills of the agents while the floor manager is planning the

[4] We only look at the departments of the primary core business activity, so that f.e. HR, Controlling, IT etc. are excluded.

service schedule and therefore controlling the capacities. 2nd level agents are promoted 1st level agents who, while still answering calls, also serve as a link between the team leader and the 1st 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. The results of the evaluation are 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 evaluated criteria are slightly different for 1st and 2nd level agents. The supervisor can rate the employee for each criterion on a scale from 1 to 5, where 5 is the highest rate and a 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 that 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 2nd level agent sitting beside the monitored agent or recorded without the agent being informed. This randomly selected call is then evaluated in a narrowly defined rating sheet and the test is passed by at least reaching $80 - 100\%$ of the score. The measure evaluating the speed of work is called the 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).

3.2 Empirical Approach and Sample Selection

3.2.1 Empirical Approach

As we observe individual ratings Y_{it} of an agent i who is evaluated at time t by an assessor $j\left(i,t\right)$ we estimate the following baseline specification:

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

Y_{it} indicates the overall subjective assessment of an agent in year t, X_i represents the main indicators for social attachment which will be explained in the following and the vector V 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 ran ordered probit regressions. To control for unobserved heterogenity in personal characteristics of the employees we also estimate individual fixed and random effects models.

At the end of the appraisal criterion catalogue the assessor was asked to give an overall impression. 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". It is important to 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)).

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 ($H1$). For every supervisor the absolute number of evaluations she conducted per year was 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 increasing time of acquaintance ($H2$). This dummy takes the value 1 for the appraisals that are "repeated" evaluations of the same supervisor and 0 otherwise.

Performance measures used as control variables are the average result of the TM, two dummies measuring the AHT performance and the standardized sum of the absence days during the period covered by a yearly subjective performance evaluation. The two dummies controlling for the AHT are conceived as follows: One of the dummies takes the value 1 for average AHT values that lie below the 90% of the mean AHT in a group per year and the other one takes the value 1 for average AHT values lying above this mean value (above the 100%) per group a year. The reason for this structure is an asymmetric evaluation of deviations from the target AHT. The company wants to save costs by having shorter calls but also provide an acceptable quality. Hence, too short calls are also sanctioned. Other control variables cover the individual-specific characteristics (age, $(\text{age})^2$, tenure and sex) and the unit-specific attributes (average age in the unit, percentage of women per unit).[5] Additionally a dummy variable is included that controls for a supervisor conducting an appraisal for the first time in 2006 or 2007.[6]

We restrict our sample to full-time employees during the years 2004−2007. Additionally we only look at 1st 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. Thus we were unable to identify the correct one.

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. Additionally we were not able to assign objective measures to every observation, 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 under 18 different supervisors. These 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.

[5] As a robustness check we also included a dummy for the sex of the supervisor and a dummy for the interaction of the same sex. Both coefficients were insignificant.

[6] Landy and Farr (1980), for instance, state that new supervisors tend to evaluate more negatively than their more senior colleagues do.

	OLS	Ordered Probit
Overall appraisal	(1)	(2)
Assessments per year	-0.00467^{***}	-0.0243^{***}
	(0.0011)	(0.0052)
TM	0.00765^{***}	0.0369^{***}
	(0.0021)	(0.0089)
Days of absence	-0.0351^{**}	-0.170^{***}
	(0.015)	(0.062)
Over AHT (Dummy)	-0.0222	-0.140
	(0.032)	(0.17)
Under 90% AHT	0.0209	0.0886
	(0.034)	(0.18)
New Assessor	-0.240^{***}	-1.028^{***}
	(0.053)	(0.22)
Female	-0.0565^{*}	-0.300
	(0.033)	(0.18)
Age	0.0241^{**}	0.126^{**}
	(0.011)	(0.063)
Age^2	-0.000293^{*}	-0.00151^{*}
	(0.00016)	(0.00086)
Av. Age Unit	yes	yes
Prop. Males Unit	yes	yes
Year Dummies	yes	yes
Constant	3.072^{***}	
	(0.41)	
Observations	520	520
R^2	0.15	
Pseudo Likelihood		-179.87569

Robust Standard errors in parentheses.

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

Table 1: Number of Assessments: OLS and Ordered Probit

3.2.2 Results

The estimation results for hypothesis 1 are reported in table 1. Column (1) shows the OLS regression results. The coefficient for the variable counting the number of assessments per supervisor-year is negative and significant at the 1%-level. With an increasing number of employees per supervisor and a resulting more distant relationship between the supervisor and her employees, subjective appraisals thus tend to be more negative. Appraisals in units with closer contact between the employees and the supervisor are thus more lenient. When comparing employees with the same transaction monitoring results, average handling times, and days of absence, those employees who work in larger teams will on average get worse assessments. We also ran an ordered probit regression confirming our result at the same levels of significance (columns (2) of table 1).

The coefficients of the objective performance measures show the expected signs and significance. High *Transaction Monitoring* results positively affect the overall assessment, while the *days of absence* have significantly negative impact. The two dummies regarding the *Average Handling Time* have no significant influence but the expected sign. The influence of being a "new assessor" has also the anticipated negative impact (significant on the 1%-Level) in the estimations.

In order to control for further unobservable heterogeneity (such as individual abilities not captured by the objective performance measures), we ran random and fixed-effects regressions.[7] The results are in line with the previous observations (table 2). Hence, 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 remains the same.

To investigate hypothesis 2 we now look at the effect of repeated assessments by the same supervisor. We therefore created a dummy variable indicating that the employee has been evaluated by the same assessor before. As repeated assessments may also capture simple tenure effects we also control for firm tenure. The results of OLS and ordered probit regressions

[7] Here we consider only employees that have been evaluated at least twice.

	Random effects	Fixed effects
Overall appraisal	(1)	(2)
Assessments per year	-0.00316^{**}	-0.00573^{**}
	(0.0014)	(0.0023)
TM	0.00536^{**}	0.00448
	(0.0024)	(0.0048)
Days of absence	-0.0447^{**}	-0.0302
	(0.018)	(0.035)
Over AHT (Dummy)	-0.0344	0.0640
	(0.046)	(0.073)
Under 90% AHT (Dummy)	0.0448	0.00281
	(0.043)	(0.078)
New Assessor	-0.224^{***}	-0.280^{**}
	(0.072)	(0.13)
Female	-0.0329	
	(0.042)	
Age	0.0322^{**}	-0.0823
	(0.015)	(0.15)
Age^2	-0.000423^{**}	-0.000279
	(0.00021)	(0.0016)
Av. Age Unit	yes	yes
Prop. Males Unit	yes	yes
Year Dummies	yes	yes
Constant	2.917^{***}	7.036^{*}
	(0.58)	(3.98)
Observations	264	264
R^2	0.19	0.13

Robust Standard errors in parentheses.

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

Table 2: Number of Assessments: Random and Fixed Effects

	OLS		Ordered Probit	
Overall appraisal	(1)	(2)	(3)	(4)
Repeated Appraisal	0.131^{***}	0.116^{***}	0.889^{***}	0.859^{***}
	(0.042)	(0.042)	(0.26)	(0.27)
Tenure	0.00870	0.0148^{*}	0.0624	0.122^{**}
	(0.0079)	(0.0080)	(0.046)	(0.053)
Female	-0.0581^{*}	-0.0620^{*}	-0.332^{*}	-0.403^{**}
	(0.034)	(0.033)	(0.19)	(0.20)
New Assessor	0.0913^{**}	-0.0767	0.787^{***}	0.749
	(0.040)	(0.057)	(0.26)	(0.57)
TM		0.00905^{***}		0.0551^{***}
		(0.0026)		(0.011)
Days of Absence		-0.0297^{*}		-0.171^{**}
		(0.016)		(0.071)
Over AHT (Dummy)		-0.0364		-0.281
		(0.031)		(0.21)
Under 90%AHT (Dummy)		-0.0152		-0.154
		(0.034)		(0.21)
Ass. Dummies	*yes*	*yes*	*yes*	*yes*
Year Dummies	*yes*	*yes*	*yes*	*yes*
Constant	2.958^{***}	2.199^{***}		
	(0.064)	(0.23)		
Observations	520	520	520	520
R^2	0.20	0.24		
Pseudo Likelihood			-164.87778	-152.14404

Robust Standard errors in parentheses.

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

Table 3: Repeated Appraisals: OLS and Ordered Probit

are reported in table 3.[8] Note that columns (1) and (3) contain results for specifications without the performance measures. The repeated appraisal dummy is significantly positive in all specifications which is well in line with our second hypothesis. Note that the inclusion of the "repeated appraisal" dummy changes the sign of the "new assessor" dummy. Hence, these results indicate that new assessors do not give worse grades because they are more strict but rather that they have not build up closer social ties with the rated employees.

Furthermore, note that the comparison of the results with and without the objective performance measures shows an interesting pattern. Due to on the job learning and human capital formation we would expect a better performance of more tenured employees. From this perspective we should expect that the inclusion of the objective performance measures would strongly reduce the size of the tenure coefficient in (2).

Interestingly, we get the opposite result as the tenure coefficient gets even stronger. This can be best understood when considering the two graphics in Figure A1 which illustrate average Transaction Monitoring scores and days of absence per year of tenure. It is striking to note that the TM results do not increase with tenure and even fall beginning with the fifth year of tenure. Furthermore, 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 seem to increase and performance may go down due to burn-out effects. In addition, there seem to be strong selection effects. Very able 1st level agents will be promoted to the 2nd level. On the other hand very poorly performing agents will leave the company. To control for these selection effects we also ran random and fixed effects estimations (table 4) with all employees who have been assessed at least twice – either by the same or by different supervisors.

The repeated appraisal dummy is still significantly positive in the random effects specifications. However, the effect is not significant in the fixed effects

[8] A comparison of the average appraisal results for the repeated assessments and the first assessments by a specific supervisor is shown in A2. With an average of 3.017 the repeated appraisals reach a marginal better score.

	Random effects		Fixed effects	
Overall appraisal	(1)	(2)	(3)	(4)
Repeated Appraisal	0.118**	0.113**	0.000693	0.0222
	(0.050)	(0.047)	(0.16)	(0.15)
Tenure	−0.00522	0.00187	−0.111	−0.103
	(0.010)	(0.011)	(0.078)	(0.076)
Female	−0.0396	−0.0331		
	(0.040)	(0.040)		
New Assessor	0.191	0.135	0.236	0.131
	(0.15)	(0.16)	(0.44)	(0.43)
TM		0.00486		0.00427
		(0.0035)		(0.0047)
Days of absence		−0.0354*		0.0101
		(0.020)		(0.038)
Over AHT (Dummy)		−0.0266		0.0953
		(0.044)		(0.069)
Under 90% AHT (Dummy)		0.0240		−0.0535
		(0.044)		(0.083)
Ass. Dummies	yes	yes	yes	yes
Year Dummies	yes	yes	yes	yes
Constant	2.965***	2.596***	3.426***	2.721***
	(0.052)	(0.30)	(0.44)	(0.43)
Observations	264	264	264	264
R^2	0.22	0.25	0.22	0.25

Robust Standard errors in parentheses.

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

Table 4: Repeated Appraisals: Random and fixed effects

specifications. This will at least partially be due to the limited number of observations with repeated assessments by the same supervisor. On the one hand, this is a strength of the data set, as it allowed us to identify the effect of team size on the evaluation of a given employee as employees have frequently moved between teams. On the other hand, this reduces the possibility of repeated assessments. In fact, only 59 appraisals out of the 264 investigated in the panel data models of table 4 have been made by the same supervisor before. Out of these 59 appraisals only $13,8\%$ were differently rated than before which means that variation in the subjective evaluation might be too low in order to observe significant effects in a Fixed Effects model.

4 Conclusion

We investigated possible distortions in subjective performance appraisals and find some evidence for the hypothesis that subjective performance is biased by social proximity of supervisor and subordinate. Our analysis shows that the size of the work unit has a negative influence on the subjective performance evaluation. After controlling for objective performance measures employees in large units received a lower evaluation than employees in smaller units. This even holds in a fixed effects regression such that the same person receives on average lower ratings when moving to a larger team.

We also observed that employees who have been evaluated by the same supervisor before receive higher ratings even with the same level of performance and the same tenure. However, the result is less robust than the previous effect when controlling for unobserved heterogeneity of the agents.

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

Baker, G., Gibbons, R. and Murphy, K. J. (1994): Subjective Performance Measures in Optimal Incentive Contracts. Quarterly Journal of Economics, 109, pp. 1125–56.

Bandiera, Oriana, Barankay, Iwan and Rasul, Imran (2008): Social capital in the workplace: Evidence on its formation and consequences. Labour Economics, 15, pp. 725–749.

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

Brandts, J. and Solà, C. (2006): Personal Relations and their Effect on Behaviour in an Organizational Setting: An Experimental Study. Working Paper, pp. 1–29.

Cardy, Robert L. and Dobbins, Gregory H. (1986): Affect and Appraisal Accuracy: Liking as an Integral Dimension in Evaluating Performance. Journal of Applied Psychology, 71(4), pp. p672 – 678.

Gibbs, M., Merchant, K. A., van der Stede, W. A. and Vargus, M. E. (2003): Determinants and Effects of Subjectivity in Incentives. The Accounting Review, 79, pp. 409–436.

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

Kane, Jeffery S., Bernardin, H. John, Villanova, Peter and Peyrefitte, Joseph (1995): Stability of Rater Leniency: Three Studies. Academy of Management Journal, 38(4), pp. 1036 – 1051.

Kingstorm, Paul O. and Mainstone, Larry E. (1985): An Investigation of the Rater-Ratee Acquaitance and Rater Bias. Academy of Management Journal, 28(3), pp. p641 – 653.

Landy, F. J. and Farr, J. L. (1980): Performance Rating. Psychological Bulletin, 87, pp. 72–107.

Lefkowitz, Joel (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), pp. p67 – 85.

MacLeod, W. Bentley (2003): Optimal contracting with subjective evaluation. America Economic Review, 93 - 1, pp. 216–240.

Milcovich, G. T. and Wigdor, K. A. (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, pp. 67–80.

Murphy, K. J. (1992): Performance Measurement and Appraisal: Motivating Managers to Identify and Reward Performance. In: Burns, W. J. Jr. (Ed.) Performance Measurement, Evaluation, and Incentives, Harvard Business School Press, Boston, MA.

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

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

Prendergast, C. J. (2002): Uncertainty and Incentives. Journal of Labor Economics, 20, pp. 115–37.

Prendergast, C. J. and Topel, R. H. (1993): Discretion and Bias in Performance Evaluation. European Economic Review, 37, pp. 355–65.

Prendergast, Canice J. (1999): The Provision of Incentives in Firms. Journal of Economic Literature, 37, pp. 7–63.

Varma, Arup, Denisi, Angelo S. and Peters, Lawrence H. (1996): Interpersonal Affect and Performance Appraisal: A field Study. Personnel Psychology, 49(2), pp. p341 – 360.

5 Appendix

Variable Group and Description	Mean	SD
Dependent Variable		
Overall assessment	2.967	0.336
Indicators for favoritism		
Assessments per year (by supervisors)	32.994	
Repeated Dummy	0.113	
Objective Performance Measures		
Result Transaction Monitoring (TM)	90.554	8.992
Over mean 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	32.216	2.425
Share of female employees	0.384	
Dummy new assessor (1/0)	0.131	

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

Table A1: Descriptive Statistics

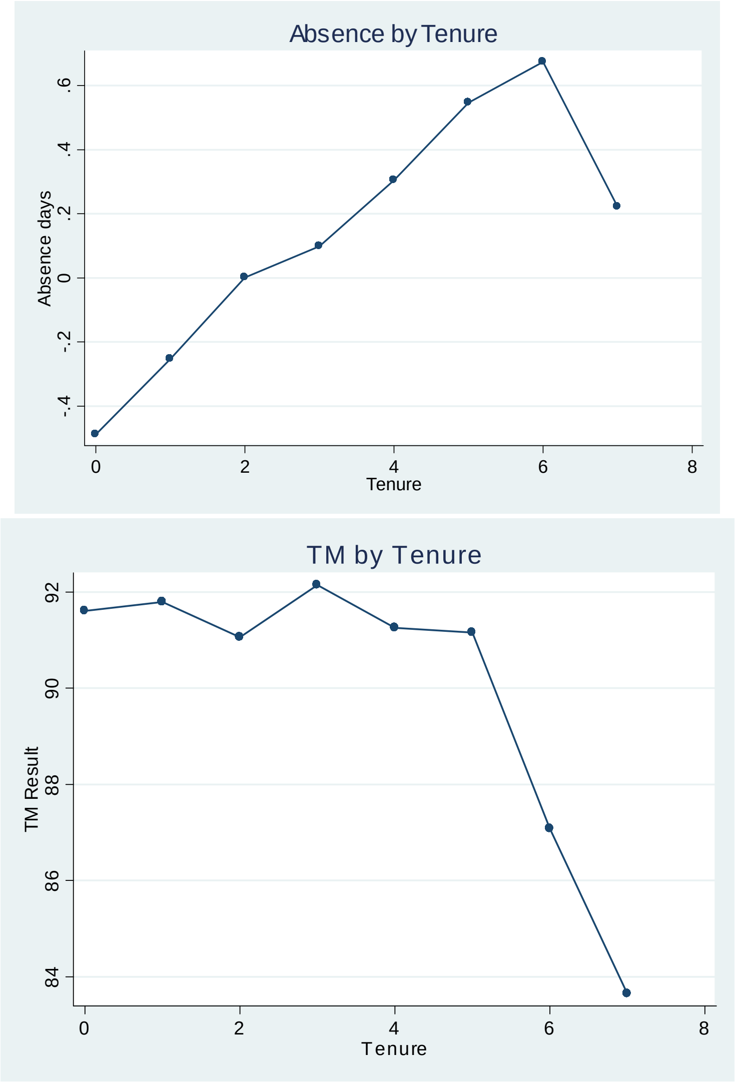

Figure A1: Performance measures by years of tenure

Repeated Assessment	Mean	Std. Dev.	Frequency
0	2.961	0.347	461
1	3.017	0.227	59

Table A2: Mean assessment of repeated/ non-repeated appraisals