

Subjective Performance Evaluations and Employee Careers

By

Anders Frederiksen (ICOA at Aarhus University, CCP and IZA)¹

Fabian Lange (Yale University, CES-Ifo, and IZA)

and

Ben Kriechel (ROA at Maastricht University and IZA)

Abstract

Supervisor ratings are common tools used by firms to evaluate employees when objective performance measures are unavailable. We examine personnel data from six large firms to determine empirical regularities in how subjective performance ratings and career outcomes are jointly distributed. Despite some differences in how firms use these ratings, we find that there are many similarities in how performance ratings are jointly distributed with key economic variables such as pay, promotions and separations. Contrary to concerns in the literature that subjective ratings only contain limited information, we find that performance ratings are important in shaping employees' careers and that they are systematically related to outcomes such as pay, job separation, and promotion.

JEL: M5

1. Introduction

How firms motivate and select employees when facing limited information about their actions and characteristics is the central issue in personnel economics. In some settings, objective performance

¹ This project would not have been possible without the exemplary willingness of a number of researchers to share their data. We thank Michael Gibbs for providing access to both the BGH and the GH data, Gerard Pfann for giving access to the Fokker data, and Lucca Flabbi and Andrea Ichino for allowing us to work on the FI data and sharing their code with us. We greatly appreciate the help we have received by all of these in working with and understanding the data. We are also indebted to the two companies referred to as FT and F for allowing us to work on their data and in particular to the employees in these companies who have made it possible. We are grateful for the comments and discussions we have had with Nikolaj Andreas Halse and Thomas Bech Hansen and members of the CCP network. Michael Lykke Jensen has provided competent research assistance on this project.

measures exist and firms rely on these objective measures to improve corporate performance.² Often, however, objective measures are unavailable. Objective measures are unlikely to exist when workers perform many different tasks in frequently changing environments, when they work in teams, and when their actions affect the value of the firm both in the short and the long run. In such contexts, employers often ask supervisors to subjectively evaluate workers' performance. Because of this subjective evaluations are a standard feature of many employment relationships and are presumably important tools in setting incentives and screening workers. Unfortunately, the empirically literature on subjective performance measures is thin which leads Oyer and Schaefer (2010) to conclude that: "we believe there is a great need for more empirical research on the use of implicit contracts and subjective performance evaluation in employment relationships".

A major obstacle in studying the use and consequences of subjective performance ratings is of course their subjectivity. Consequently, subjective evaluations can be influenced not just by the performance of the worker, but also by the characteristics of the supervisor and by the relation between the supervisor and the employee. Furthermore, subjective measures are reported on arbitrary metrics. This implies that, the ways abstract concepts such as effort and ability - concepts common in models of moral hazard and adverse selection - map into performance ratings can potentially vary widely across firms and circumstances. The consequence is that it is potentially difficult to evaluate the standard models in personnel economics using subjective performance ratings. The question arises if subjective ratings do contain any useful information for empirical research in this field. The answer to this question depends in part on whether there are empirical regularities across firms in how subjective ratings co-move with objective career outcomes.

In this paper, we answer this empirical question by studying multiple firm-level personnel data sets containing subjective performance ratings. We search for regularities in how subjective performance measures are related to a wide set of career outcomes that spans wages, bonus pay, total compensation, demotions, promotions, and separations (sometimes distinguished by dismissals and quits). We examine how performance ratings and outcomes correlate both within and across

² Over the last decades, a number of studies have examined the use of objective performance measures. In the most famous of these studies, Lazear (2000) explored how Safelite, a windshield repair-company, used the number of windshields replaced as a measure of performance for its repairmen. This landmark paper showed how a change in the incentive system affected both the sorting and selection of workers and their performance on the job. Other objective performance measures that have been investigated are the number of trees planted (Shearer (2004)) and the amount of fruit picked (Bandiera, et al. (2005, 2007)). These studies considered how objective performance varies with the pay system. In addition to this a burgeoning literature in education economics uses value added measures of student test scores to examine issues in selecting and incentivizing teachers (Barlevy and Neal (2011), Goldhaber and Hansen, 2010).

periods and how the relations between ratings and outcomes change with tenure and experience. Some features of the data seem to be unique to individual firms and are likely to be caused by specificities in the rating systems and human resource practices used by different firms. But, other patterns are sufficiently consistent across firms that we cautiously propose them as being general empirical regularities.

Our analysis is based on data from six companies for which we have been able to obtain personnel data that includes subjective performance measures.³ These six data-sets have been analyzed by us and by other researchers in other contexts before. Among them, the most prominent is the one analyzed by Baker, Gibbs, and Holmström in a series of articles on internal labor markets (Baker, Gibbs and Holmström, 1993, 1994a,b). These papers have inspired important theoretical contributions in personnel economics (e.g. Gibbons and Waldman, 1999, 2006). More recently, one of us has exploited this data to study the dynamics of individual productivity and employer learning (Kahn and Lange (2011)). Others (DeVaro and Waldman (2011)) used this data to test the promotion-signaling hypothesis. A second data-set which also originates from the US has been used by Gibbs and Hendricks (2004) to examine the role of formal salary systems. The other data-sets are from Europe. We have data from a large Italian bank which Flabbi and Ichino (2001) used to replicate and expand upon the analysis of Medoff and Abraham (1980, 1981). We also examine data from Fokker, a now defunct Dutch aircraft manufacturer whose human resource practices are the subject of Dohmen (2004) and Dohmen, Kriechel, and Pfann (2004). Another data-set comes from a large pharmaceutical company. Frederiksen and Takáts (2011) used this data to study the mix and hierarchy of incentives. These data did not include subjective performance evaluations, but we were able to obtain a second wave that includes supervisor ratings. The last of our data-sets was used by Frederiksen (2010) to analyze explicit and implicit incentives in a large bank.

There are several common patterns in these six personnel data-sets. To begin, we find that the support of the rankings is highly restricted in all firms. With one exception, the companies use a five point performance scale. The effective support of the ratings is restricted further because supervisors are reluctant to give bad ratings; there is clearly a “Lake Wobegon”-effect. Typically, more than 95 percent of ratings are concentrated on only three values at the upper end of the ranking scale.

³ We are not aware or have been unable to access other data-sets that contain these measures. Most notably, the data used in Medoff and Abraham (1980, 1981) is unfortunately not accessible.

A second finding common to all firms is that experience and tenure fail to explain much of the variation in performance evaluations. Job levels, however, explain a fairly large component of the variation in performance ratings. We also observe that the gradient of performance rankings in experience and tenure (controlling for job level) varies greatly across firms. In some firms, rankings decline with experience and tenure, whereas in other firms they increase. In contrast, earnings gradients are quite similar across firms and they are very robust to controlling for performance rankings. This reflects the fact that evaluations and experience/tenure correlate only weakly.

Medoff and Abraham (1980, 1981) interpret the fact that earnings gradients are robust to controlling for performance rankings as evidence that earnings gradients with experience do not reflect average productivity difference within job grade. The empirical evidence in favor of this argument relies on assuming that performance ratings are fairly accurate measures of productivity and that they can be used to compare productivity across experience levels. We are not willing to make these assumptions. Rather, we believe that the data is better explained by assuming that performance ratings are relative measures that rank workers with similar levels of experience. And, we believe that these performance rankings contain a significant amount of noise. The first assumptions free up the experience profiles of performance rankings, since they are taken to be simple ordinal rankings within experience levels. Furthermore, the assumption that there is significant noise can explain why earnings profiles are fairly robust to controlling for performance ratings. We thus assume that performance ratings reflect relative performance within a peer group defined by demographics, education, and experience. Consequently, in the remainder of this paper, we interpret performance rankings as ordinal rankings of workers within a narrowly defined peer group rather than a cardinal ranking of performance as proposed by Abraham and Medoff (1980, 1981). The practical consequence is that the remainder of the analysis (unless stated otherwise) is based on the residuals obtained from regressions of performance rankings that control for flexible functional forms in demographics, education, experience, and calendar time.

We then turn to the correlations of the (orthogonalized) performance rankings across experience. We find, without exception, that these idiosyncratic components are highly correlated at short lags. At one lag, the autocorrelations almost always exceed 0.4, typically exceed 0.6, and sometimes exceed 0.8. The autocorrelations decline with longer lags and tend to be between 0.1 and 0.4 after three or four lags. The autocorrelations in performance evaluations are also found to be higher for more experienced workers.

Of key interest is how idiosyncratic components of performance and pay correlate. Performance evaluations are positively correlated with total compensation and also with base pay and bonus. Typically, we find that log base pay correlates more highly with contemporaneous and past performance evaluations than with performance evaluations from the future and the correlation of base pay with performance evaluations is higher for more experienced than less experienced workers. These are the patterns that have been exploited by Kahn and Lange (2011) in their paper on employer learning and heterogeneous productivity changes. Regarding bonuses, we find that in the correlations of performance rankings with bonuses are always positive. However, in some firms log bonuses correlate very highly with current performance evaluations and less so with future or past performance rankings. This pattern is what we would expect if firms tie bonuses directly to current performance. In other firms, however, there is little difference in how bonuses correlate with current, past, or future performance rankings. Finally, because base pay makes up a substantially larger component of total compensation than bonuses, correlations between performance and total compensation largely resembles those between performance and base pay.

Performance ratings influence how employees move up the firm's hierarchy. Promotion probabilities vary considerably across firms, in part because these firms differ in their organizational design. The ratio of promotions to demotions ranges between 3.1 and 80 (in one firm no demotions are observed) but half the firms have promotion/demotion ratios less than 4.6. Thus, demotions are not uncommon, even though they are less frequent than promotions. In all firms, promotions correlate positively and demotions are correlate negatively with performance.

Transitions out of the firm are negatively correlated with performance. In the two firms where we can distinguish dismissals from quits we find that both dismissals and quits are negatively correlated with performance rankings, and that the correlation between performance and dismissals is larger.

Our analysis of the six firm-level datasets proceeds as follows. We introduce the firms and present simple descriptive statistics on subjective performance evaluations in the next section. Section 3 is inspired by the work of Medoff and Abraham (1980, 1981) and considers how subjective performance ratings vary with experience and tenure. In this Section, we take a stand on how to interpret subjective performance evaluations. In Sections 4 and 5, we analyze the autocorrelation patterns of performance and pay respectively. In Section 6, we establish how total compensation as well as its components - base pay and bonuses – are related to performance ratings. Sections 7 and 8

address the importance of subjective performance evaluations for employee mobility both internally (promotions and demotions) and out of the firm (separations, quits and dismissals). We conclude in Section 9.

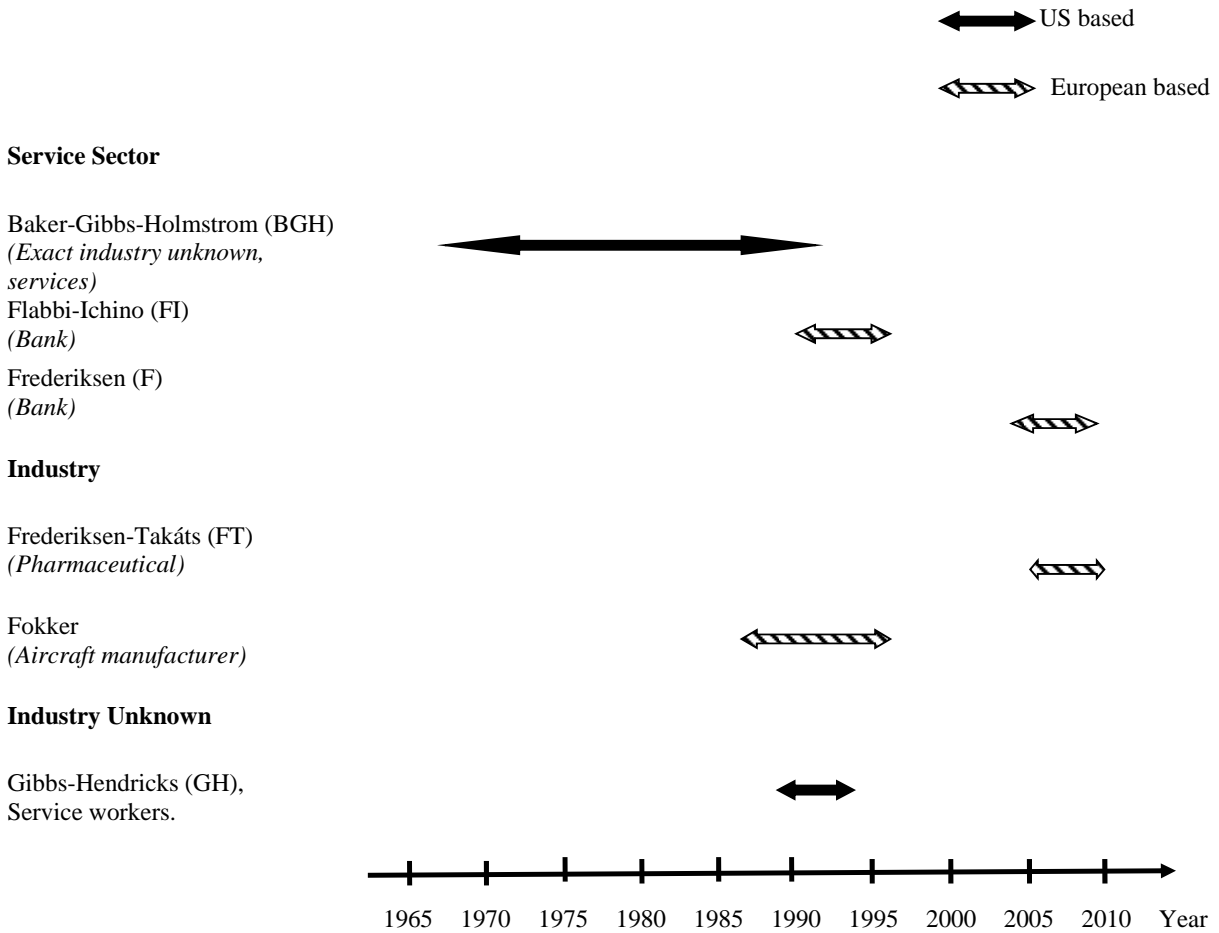
2. The Firms

We study personnel data from six large companies that are very diverse. On one qualitative aspect, however, the datasets are similar. That is, with the exception of Fokker, a Dutch airplane manufacturer, and the pharmaceutical company referred to as FT who have data on blue-collar and white-collar workers alike, all companies cover only white-collar workers. Otherwise, the six companies differ in terms of where they are located, what industries they operate in, and during what time-period the data was collected. Figure 1 summarizes how the firms differ according to these criteria. The Baker-Gibbs-Holmstrom (BGH, hereafter) and the Gibbs-Hendricks (GH) companies are based in the US whereas the remaining four companies (Flabbi-Ichino (FI), Frederiksen-Takáts (FT), Frederiksen (F) and Fokker) are based in Europe.⁴ The companies span several sectors. BGH is in the service sector.⁵ FI and F operate in the financial sector. FT is a pharmaceutical company and Fokker was an aircraft manufacturer. We do not know what industry GH belongs to. The companies also span different time-periods. BGH covers the period 1969-1988 and thus provides the earliest data available. FI, GH, and Fokker provide data from the late 1980s up until the mid-1990s. The most recent data stem from FT and F and cover the period from the early 2000s to 2010. Thus, the companies are operating in important markets in Europe and the US, they cover several industries and span the time period from the late sixties to the present time. Because of these differences it is expected that the use of performance evaluations vary largely across the companies. If, however, consistent patterns across the firms emerge they can be interpreted as general empirical regularities. But, before such regularities are established we present the firms.

⁴ FI is located in Italy and Fokker operated out of the Netherlands until it went out of business in 1996. FT and F are still in operation and for this reason their precise location and identity remain unavailable.

⁵ We are restricted from revealing the exact sector.

Figure 1. Location, Industry and Time Period



Baker-Gibbs-Holmstrom (BGH)

Two ground-breaking papers (Baker, Gibbs and Holmstrom (1994a, b)) analyzed the personnel data of a US based service sector firm. The data contains information on managerial employees (about 20% of the workforce) and covers the period 1969 to 1988 which was a period where the firm experienced rapid growth in assets and employees. BGH described the internal personnel structure of this firm in detail and they considered particularly the question whether the data supports or refutes the notion of an ‘internal labor market’. They also considered in an informal way if the data was consistent with models of employer learning, human capital acquisition, and simple provision of incentives. In summarizing the findings of BGH, Gibbs (1994) writes that BGH “concluded that

their evidence was inconsistent with simple models of learning and incentives. Instead, they suggested that many of their findings were consistent with a model in which employees accumulate human capital at varying rates.”

BGH did not analyze the use of subjective performance ratings in this firm. The first analysis of the use of subjective ratings in this firm comes from Gibbs (1994). He showed that performance ratings correlated strongly with pay, pay rises, and promotions, but they did not predict exit from the firm. Similar to BGH and based on the same data, Kahn and Lange (2011) reestablish that heterogeneous human capital accumulation is important, but by using the information conveyed in the subjective ratings they also provide evidence for the fact that employer learning is taking place at all stages of the employees’ careers: That is, employers are trying to ‘hit a moving target’. Our analysis of the firm replicates the results in Gibbs (1994), but it also goes beyond his analysis by focusing more on dynamic aspects. Further, the dynamic analysis presented in Section 6 is designed to test if the core results established in Kahn and Lange (2011) are present in other firms as well.

Three peculiarities of BGH are particularly worth mentioning. First, no variable in the original data explicitly identifies the job hierarchy. Instead, BGH used the internal mobility patterns and some information on job titles to deduce the hierarchy. In our analysis we will rely on the hierarchy identified by BGH in their original work. Second, while compensation is broken down into base pay and bonuses for the years 1981 to 1988, we only have data on base pay prior to 1981. Bonuses make up a small fraction of total compensation in the later years and for this reason we use the compensation data from the entire 1969 to 1988 period for our work on pay. When we look specifically at bonuses and base pay, we restrict the data to the years where the two types of income are available separately. Third, tenure data can only be calculated precisely for workers entering after 1969, when the sample period starts. Any statistics related to tenure that we present below are based on those observations for which tenure is available. By contrast, experience is measured as potential experience ($\text{age} - 6 - \text{years of schooling}$). We will use this measure of experience in the analysis of all data-sets.

As shown in Table 1, BGH consists of 55,754 employee-year observations from a total of 9,747 unique employees.⁶ Average total compensation (in 2000 dollars) is about 80,000, which far

⁶ In our analysis of the firms we only use employees with experience less than 40.

exceeds the average for the US population.⁷ This, as well as the demographics and the high education level reflect the fact that the BGH data consists of managerial employees.

Gibbs-Hendricks (GH)

Our description of GH is based on Gibbs and Hendricks (2004). The data covers white-collar professional and managerial employees as well as clerical and technical office workers employed in a large US corporation active in several different businesses. The industry is unknown to us. The data, which includes tenure as well as data on compensation, promotions and demotions spans the years 1989 to 1993.

Interestingly, GH contains information on the administrative pay system used to set pay (Grade, Hay, and PAQ – described in detail in GH). Gibbs and Hendricks report that the “nominally different systems, Hay, Grade, and PAQ, in practice were very similar in design”. These systems assigned target salary ranges for different jobs in the firm and then positioned individuals within these ranges using the individuals’ past location in the range as well as their performance rankings. A major question asked by Gibbs and Hendricks is to what extent these administrative rules simply reflect market forces (they act as a “veil”). Their findings related to this question are somewhat mixed. For instance, while they find that zero raises are often due to individuals bumping up to the top of their salary ranges, they also find that supervisors seem to be using bonuses to partially alleviate these constraints. These constraints can also be avoided by promoting workers to higher salary ranges. Overall, they argue that the firm does not incur large costs from the nominal constraints imposed by the formal salary rules. This is consistent with the view that assignment rules to jobs combined with bonuses and some discretion in pay is sufficient to accommodate market forces.

GH contributes a total of 43,964 employee-year observations from a total of 14,372 unique employees who all have less than 40 years of experience. Like BGH but less dramatically so, their average compensation of \$58,000 exceeds the US average. The data contains indicators for promotions and demotions but we cannot distinguish between those in management positions and those not.

⁷ All earnings measures presented in this paper are reported in 2000-dollars equivalents.

Fokker

Fokker was a Dutch Airplane Manufacturer that got into financial trouble after 1991 and underwent several rounds of downsizing before finally going bankrupt in 1996. The data we are analyzing covers the years 1987-1996. This data consists of both blue-collar and white-collar workers who were subject to very different personnel regimes. We therefore analyze the blue-collar and white-collar samples separately. If employees are represented in both groups at different points in time they were dropped from the analysis. There are 71,086 employee-year observations in our blue collar data and 11,516 unique blue collar workers. The white collar sample is smaller and has 25,771 employee-year observation and 4,102 unique individuals.

The performance ratings in this firm were tied to compensation according to a very strict system of rules and regulations. As we will see below, these rules are reflected in the correlation patterns between compensation and performance ratings, particularly among blue-collar workers. For a more detailed description of this data see Dohmen (2004) and Dohmen et al. (2004).

Flabbi – Ichino (FI)

The company referred to as FI is a large bank operating through the Italian peninsula. Flabbi and Ichino (2001) use this data to replicate the analysis by Medoff and Abraham (1980, 1981) and find results that are consistent with those earlier studies.

A detailed description of this data is given in Flabbi and Ichino (2001) and we follow them in most aspects when constructing the dataset. Over a period of 6 years from 1990 to 1995, 12,996 unique employees contributed with a total of 63,390 employee-year observations.

Subjective performance evaluations are available only for non-managerial workers. As do Flabbi and Ichino, we restrict the sample to male employees. Reflecting the lower incomes in Italy and the restriction to non-managerial employees, we find that average earnings are much lower than in the GH and BGH data and amount to only about \$29,000.

Fredriksen – Takáts (FT)

The FT company is a global pharmaceutical company with headquarter in Europe but production and sales activities on all continents. Frederiksen and Takáts (2011) study the firm's use of incentives and derive a hierarchy of incentives. In particular, they explain why firms often use a complex mix of incentives. That is, in the Frederiksen-Takáts model firms concerned about

employee quality may find it optimal to combine cost-effective incentives such as promotions and bonuses with dismissals. The reason is that even though dismissals are costly they (like promotions) provide both incentives and contribute to the sorting and selection of employees. Subsequently, Frederiksen and Takáts investigated the consequences of promotions and dismissals for the employees' careers and used the FT data to provide support for these predictions.

The data available for analysis contain employees working in the country where the Headquarter is located. The activities taking place in this country are extensive and cover besides headquarter activities a broad set of functions including production, IT and R&D. The data used in the analysis span the years 2006 to 2010 and thus constitutes the most recent data among the six datasets.

The use of a systematic and companywide performance appraisal system is relatively new to the FT firm and the sample period overlaps with the phasing-in of the performance measurement system. Consequently only a fraction of employees received performance ratings in the early years. But, by the end of the sample period more than two-thirds of employees received a rating. Besides this, the FT data contains all relevant information on compensation and employee mobility and a unique feature of the data is that separations can be split into quits and dismissals.

A total of 53,544 employee-year observations are available for analysis and these are based on information from 17,354 unique individuals. The wage level in this firm is \$46,000.

Frederiksen (F)

The F firm is a large bank and its provision of implicit and explicit incentives was studied by Frederiksen (2010). Frederiksen established that earnings growth heterogeneity is significant among newly recruited employees and that this heterogeneity to a large extent is driven by differences in performance ratings. Because of the detailed information on job functions available in the data it is also established that the relation between performance ratings and earnings growth differs across employee subgroups in the firm. For instance, employees in "market functions" have income progression which is highly sensitive to individual performance ratings, whereas the relationship between performance ratings and earnings growth is less pronounced in other areas of the firm.

The F firm has some international activities but our data cover only domestic operations. The data comprises more than 20,000 unique employees and a total of 89,508 employee-year observations over the years 2004 to 2009. Average earnings in the firm are close to \$ 50,000. The F data

contains all relevant information on performance, compensation and mobility and in addition to this a unique feature of the data is that separations can be split into quits and dismissals.

Table 1. Descriptive Statistics

	BGH ³	GH	Fokker Blue- Collar	Fokker White- Collar	FI ⁴	FT	F
Unique Employees	9,747	14,372	11,516	4,102	12,996	17,354	20,183
Observations	55,754	43,964	71,086	25,771	63,390	59,544	89,508
Observations with performance ratings	36,428	36,337	70,851	25,731	62,428	22,350	64,550
Fraction Managers <i>Compensation</i> ^{1,2}	Only Managers	Breakdown not clear	Na	Na	Only Non- Managers	0.098	0.260
All employees	Na	57,943 (37,055)	21,800 (4,103)	40,086 (12,851)	Na	46,014 (22,691)	48,334 (35,154)
Managers	80,069 (43,536)	Na	Na	Na	Na	69,776 (37,204)	60,930 (55,211)
Non-managers	Na	Na	Na	Na	29,128 (5,462)	43,418 (18,750)	43,738 (22,261)

1) Reported are averages with standard deviations in parentheses obtained using workers with less than 40 years of labor market experience in the respective firms.

2) All earnings are in USD 2000-prices. US data is deflated using the CPI-U. For the other data-sets we use appropriate deflation indices and convert to USD using year-end 2000 exchange rates.

3) The BGH data contains only managerial employees comprising about 20% of the total workforce. In GH and FI, the breakdown of the workforce into managerial and non-managerial employees is not clear from the information provided.

4) FI data are available from 1975 to 1995 but performance data is only available from 1990. The statistics reported are based on the period 1990 to 1995.

Subjective Performance Measures

Table 2 contains information on the performance scales and distributions used by the companies. With the exception of GH, the scale of the performance measures and their distributions are very similar. Most common is a 5-point scale arranged such that 1 corresponds to a low rating and 5 to a high rating. But, deviations from the 5-point scale are observed. For instance, Fokker applied a 5-point scale for its white collar workers and a 6-point scale for its blue collar workers. The only firm applying a substantially different scale is GH where the scale has 18 levels.

Table 2. The Distribution of Subjective Performance Measure

		BGH	GH ¹	Fokker Blue Collar	Fokker White Collar	FI	FT	F
Rating scale		1-5	18 levels, but 93% on 6 levels.	1-6	1-5	2-6	1-5	1-5
Low	1	0.05	25	0.12	0.23	.	0.13	0.13
	2	0.74	18	1.35	3.96	0.06	3.16	2.58
	3	17.05	4	43.83	81.33	2.59	49.49	42.21
	4	50.00	16	40.53	14.13	14.37	39.98	47.38
	5	32.16	24	12.70	0.35	38.01	7.24	7.70
High	6	.	6	1.48	.	44.97	.	.

Notes: 1) GH applies a 1 to 18 scale but 6 levels account for 93 per cent of the ratings. For GH, this table shows only the rates pertaining to the 6 most common ratings.

In all firms, performance ratings are concentrated on a subset of the scale. The concentration is most extreme for Fokker white collar, where one category accounts for 81% of rankings. For the other firms, typically all but 3-4% of ratings are concentrated in only 3 categories. From the distributions it is clear that managers are very reluctant to give employees low ratings as these are rarely used.

The clear majority of employees are subject to performance appraisals each year. In some cases, however, an employee subgroup is exempted from evaluations. For instance, in FT systematic performance evaluation is relatively new and during the phase-in period various employee

subgroups were exempted from the evaluation program. In other companies newly recruited employees are unlikely to obtain performance evaluations. For example, in F employees do not receive ratings in the first employment year. It is likely that similar rules are in place in other firms. In any case, the incidence of performance evaluations is not uniform and varies for reasons that are not well understood.⁸ In what follows, we treat the incidence of evaluations as exogenous.

3. Performance Ratings over the Life-cycle – Medoff and Abraham revisited

We begin our analysis of subjective performance ratings⁹ by investigating how they are related to experience and tenure. In two well-known papers, Medoff and Abraham (1980, 1981) used personnel records containing subjective performance ratings from 3 different firms to answer the challenge raised by Mincer (1974): whether it can be “shown that growth of earnings under seniority provisions is largely independent of productivity growth.”¹⁰ In their data, (i) performance measures decline with experience, *holding grade level constant*, and (ii) controlling for performance ratings did not attenuate the observed earnings-experience gradient.¹¹ Thus, because they interpret the subjective performance measures as cardinal measures of productivity they conclude that “... the primary findings ... appear to be at odds with what would be expected given the human capital interpretation of the experience earnings profile.”

In tables 3, 4, and 5, we provide evidence on the same question. Table 3 shows that there is no consistent pattern across firms in how mean performance ratings vary with experience, age, and tenure. Performance ratings increase with age, tenure, and experience in FI, they follow an inverted u-shape in GH, FT and F, and decline in BGH. Within Fokker, performance ratings increase for blue collar workers. Among white collar workers performance ratings are almost perfectly flat.

⁸ Halse et. al. (2011) study the use of performance measures in a global company and discuss why performance evaluations may differ in terms of quality and prevalence across countries.

⁹ In this Section, we use the raw performance ratings prior to orthogonalizing as described in the introduction.

¹⁰ P.80, Mincer (1974).

¹¹ Using the omitted variable bias formula, it should be clear that both of these findings are directly related in that controlling for performance ratings will attenuate the earnings-experience gradient if (i) performance ratings correlate positively with experience and (ii) performance ratings correlate positively with wages.

Table 3. Average Performance by Age, Experience and Tenure

	BGH	GH	Fokker Blue Collar	Fokker White Collar	FI	FT	F
Rating scale	1-5	2-15	1-6	1-5	2-6	1-5	1-5
Age:							
- 30	4.35 (0.64)	8.86 (1.82)	3.42 (0.59)	3.09 (0.37)	4.74 (0.76)	3.49 (0.68)	3.40 (0.64)
31 – 40	4.20 (0.69)	9.26 (1.91)	3.79 (0.76)	3.10 (0.463)	5.26 (0.75)	3.55 (0.68)	3.68 (0.69)
41 – 50	4.02 (0.73)	9.24 (1.96)	4.00 (0.83)	3.12 (0.49)	5.44 (0.74)	3.51 (0.69)	3.66 (0.67)
51+	3.90 (0.72)	9.13 (1.93)	4.29 (0.91)	3.11 (0.51)	5.58 (0.70)	3.41 (0.66)	3.56 (0.66)
Experience:							
1-10	4.33 (0.66)	8.98 (1.84)	3.38 (0.57)	3.10 (0.37)	4.76 (0.74)	3.54 (0.68)	3.42 (0.65)
11-20	4.17 (0.69)	9.26 (1.94)	3.69 (0.73)	3.10 (0.42)	5.22 (0.77)	3.54 (0.69)	3.69 (0.69)
21-30	4.00 (0.73)	9.20 (1.95)	3.97 (0.81)	3.11 (0.48)	5.43 (0.73)	3.51 (0.68)	3.65 (0.67)
31-40	3.83 (0.74)	9.08 (1.90)	4.24 (0.90)	3.11 (0.51)	5.59 (0.67)	3.39 (0.66)	3.55 (0.66)
Tenure:							
0-5	4.18 (0.70)	8.87 (1.85)	3.35 (0.57)	3.14 (0.50)	4.66 (0.74)	3.49 (0.68)	3.47 (0.70)
6-10	4.05 (0.71)	9.34 (1.92)	3.66 (0.70)	3.11 (0.46)	5.15 (0.75)	3.53 (0.68)	3.65 (0.67)
11-20	3.97 (0.77)	9.36 (1.95)	3.94 (0.77)	3.12 (0.43)	5.35 (0.75)	3.54 (0.68)	3.70 (0.66)
21+	Na	9.18 (1.92)	4.38 (0.86)	3.08 (0.40)	5.59 (0.68)	3.48 (0.69)	3.60 (0.66)

Note: Experience refers to potential experience calculated as: age – 6 – years of education. For BGH, tenure is only available for individuals entering the sample after 1969 and the tenure statistics are therefore limited to the sample of those individuals.

Table 4 presents regression results similar to those of Medoff and Abraham (1981). That is, we regress performance ratings on a polynomial in experience, a polynomial in tenure and controls. Among the controls are year and education dummies, gender, age and race when appropriate. We orthogonalize tenure using experience and the other controls so that the experience coefficients include any effect that operates through tenure. The tenure coefficients can be interpreted as “within experience” effects of tenure.

As in Table 3, we find that the performance-experience profiles are not stable across firms. At the average level of experience, we find that performance ratings decline for BGH, FT, and F and increase for GH, blue collar Fokker and FI. The shapes of the quadratic polynomials are more regular; in all firms except BGH and white-collar Fokker we find that the shapes of the quadratic experience and tenure profiles are concave. This implies that performance ratings increase more rapidly among newly hired workers and/or among young workers.

We generally find that job level indicators explain significant fractions of the variation in performance. In BGH, FI, FT, and F, job level indicators nearly double the R-square. In addition, the estimated performance gradients in experience and tenure are typically quite sensitive to controlling for job levels. In FI, F, and FT, controlling for job levels attenuates the effect of experience on performance rankings by a third to more than one-half.

It should be noted, however, that in these performance regressions R-squares are generally low and the standard errors of the regressions are large, indicating that there is substantial variation in performance that does not correlate with either experience or tenure. One explanation is that the performance rankings are noisy measures of actual productivity and that the measurement error attenuates the estimated coefficients on performance rankings in wage regressions. If this is true, then Abraham and Medoff’s finding that earnings-experience and earnings-tenure profiles are insensitive to including subjective performance rankings is plausibly due to attenuation bias.

Table 4. Experience and Tenure Profiles of Performance Ratings

	BGH¹		GH²		Fokker Blue Collar		Fokker White Collar		FI		FT		F	
Rating Scale	1-5		2-15		1-6		1-5		2-6		1-5		1-5	
Experience	-0.013 (0.002)	-0.035 (0.002)	0.071 (0.004)		0.050 (0.001)	0.050 (0.001)	0.002 (0.001)	-0.005 (0.001)	0.070 (0.002)	0.034 (0.002)	0.013 (0.003)	0.005 (0.003)	0.038 (0.001)	0.015 (0.001)
Experience squared / 100	-0.011 (0.004)	0.028 (0.004)	-0.162 (0.011)		-0.045 (0.003)	-0.045 (0.003)	-0.005 (0.003)	0.006 (0.004)	-0.093 (0.003)	-0.043 (0.004)	-0.041 (0.006)	-0.026 (0.006)	-0.079 (0.003)	-0.029 (0.003)
Orth. Tenure	-0.034 (0.003)	-0.095 (0.004)	0.101 (0.005)	Na	0.058 (0.001)	0.059 (0.001)	0.012 (0.001)	0.010 (0.001)	0.078 (0.002)	0.052 (0.242)	0.013 (0.002)	0.008 (0.002)	0.014 (0.001)	0.014 (0.001)
Orth. Tenure squared / 100	0.285 (0.024)	0.489 (0.024)	-0.322 (0.020)		-0.081 (0.004)	-0.082 (0.004)	-0.013 (0.004)	-0.010 (0.004)	-0.157 (0.006)	-0.129 (0.006)	-0.026 (0.007)	-0.014 (0.007)	-0.027 (0.003)	-0.024 (0.002)
Job level controls	NO	YES	NO		NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Experience effect at the mean	-0.016	-0.025	0.019	Na	0.037	0.037	0.000	-0.003	0.033	0.017	-0.144	-0.172	-0.132	-0.026
R-squared	0.09	0.17	0.04	Na	0.23	0.23	0.01	0.02	0.14	0.24	0.02	0.05	0.07	0.14
Reg. std. Error	0.68	0.65	1.89		0.65	0.65	0.41	0.41	0.73	0.69	0.68	0.67	0.64	0.62
N	36,290	36,290	36,316		54,761	54,761	20,737	20,737	62,428	62,428	22,350	22,350	54,793	54,793

Note: Experience refers to potential experience defined as: Age – 6 – years of schooling. In each column, we residualize tenure and tenure-squared using all other controls appearing in that regression. Each regression controls for education in a flexible manner, where the exact education controls depend on the data set used. In addition to education all regressions control for gender and year as well as race dummies when appropriate.

1) In BGH, tenure is not available for those already in the firm in 1969. We substituted a value of 0 for the orthogonalized tenure measure for those with missing tenure.

2) GH does not have data on the hierarchical structure of the firm.

In Table 5, we present log earnings regression analogous to Medoff and Abraham (1980, 1981). Medoff and Abraham examined whether log earnings gradients in experience and tenure attenuate when performance ratings are included among the controls.¹² Flabbi and Ichino (2001) replicated these regressions using the FI data. We consider the same specification for log earnings used in those papers. As do Abraham and Medoff (and FI), we find only weak evidence that controlling for performance evaluations reduces the magnitude of the experience and tenure effects on earnings. Following the interpretation of Medoff and Abraham (1980, 1981) that performance ratings can be used to compare performance across individuals at different experience levels, one would be forced to conclude that average earnings differences with experience do not reflect worker productivity. We prefer a different interpretation of the performance measures.

The results in tables 3 and 4 show that experience and tenure profiles in performance ratings vary considerably across companies even when job levels are controlled for. Following Medoff and Abraham (1980), these results would implausibly imply very large differences across firms in how the productivity of workers varies with experience. We believe that the data is better explained if performance ratings are interpreted as “noisy” measures of relative performance within narrowly defined peer groups - where the peer group is defined by demographics, education and experience. The main advantage of this interpretation is that it frees up the interpretation of experience profiles of performance rankings, since ratings are taken to be simple ordinal rankings within experience levels. As a consequence of our interpretation we will, in the remainder of the paper assume that performance rankings reflect ordinal measures of performance within narrowly defined peer groups.

The next step in our analysis is to examine the data for consistent patterns in the joint distribution of performance ratings and career outcomes. We begin with the autocorrelation patterns in performance ratings (Section 4) and in the subsequent sections we investigate in detail how performance ratings and career outcomes correlate.

¹² Medoff and Abraham control for job levels in their regressions.

Table 5. Log-Earnings Functions with Pay-grades and Performance Ratings

Panel A: BGH, GH, Fokker												
	BGH ¹			GH ²			Fokker: Blue collar			Fokker: White collar		
Experience	0.037 (0.001)	0.010 (0.006)	0.012 (0.001)	0.049 (0.001)	0.045 (0.001)		0.050 (0.000)	0.046 (0.000)	0.044 (0.000)	0.062 (0.001)	0.039 (0.000)	0.039 (0.000)
Experience squared / 100	-0.070 (0.002)	-0.020 (0.001)	-0.022 (0.001)	-0.092 (0.001)	-0.085 (0.002)		-0.092 (0.000)	-0.086 (0.000)	-0.084 (0.000)	-0.094 (0.001)	-0.057 (0.000)	-0.058 (0.000)
Orth. tenure	0.054 (0.002)	0.004 (0.001)	-0.001 (0.001)	0.039 (0.001)	0.036 (0.001)		0.013 (0.000)	0.011 (0.000)	0.010 (0.000)	0.015 (0.001)	0.010 (0.000)	0.009 (0.000)
Orth. tenure squared / 100	-0.144 (0.012)	0.027 (0.008)	0.011 (0.008)	-0.097 (0.003)	-0.085 (0.003)		-0.019 (0.000)	-0.018 (0.000)	-0.015 (0.000)	-0.003 (0.002)	-0.023 (0.001)	-0.022 (0.001)
<i>Performance rating:</i>						Na						
1			Omitted		Omitted				Omitted			Omitted
2			-0.001 (0.195)		-0.056 (0.005)				0.010 (0.012)			-0.041 (0.017)
3			0.091 (0.194)		-0.048 (0.009)				0.030 (0.012)			0.003 (0.017)
4			0.114 (0.194)		0.063 (0.005)				0.073 (0.012)			0.056 (0.017)
5			0.165 (0.194)		0.095 (0.005)				0.106 (0.012)			0.115 (0.022)
6					0.137 (0.008)				0.154 (0.013)			
Job level effects	NO	YES	YES	NO	NO	YES	NO	YES	YES	NO	YES	YES
R-square	0.394	0.737	0.742	0.293	0.626	Na	0.79	0.83	0.84	0.67	0.87	0.88
N	21,474	21,474	21,474	36,316	36,316	Na	54,761	54,761	54,761	20,737	20,737	20,737

Panel B: FI, FT, F									
	FI			FT			F		
Experience	0.016 (0.000)	0.001 (0.000)	0.001 (0.000)	0.065 (0.003)	0.064 (0.003)	0.064 (0.003)	0.034 (0.001)	0.004 (0.000)	0.003 (0.000)
Experience squared / 100	-0.009 (0.000)	0.010 (0.000)	0.011 (0.000)	-0.108 (0.007)	-0.103 (0.006)	-0.100 (0.006)	-0.071 (0.001)	-0.007 (0.001)	-0.006 (0.001)
Orth. tenure	0.025 (0.000)	0.025 (0.032)	0.009 (0.000)	0.079 (0.003)	0.080 (0.002)	0.079 (0.002)	-0.001 (0.000)	-0.001 (0.000)	-0.001 (0.000)
Orth. tenure squared / 100	-0.030 (0.001)	-0.009 (0.000)	-0.001 (0.000)	-0.183 (0.009)	-0.172 (0.007)	-0.170 (0.007)	-0.000 (0.001)	0.001 (0.001)	0.002 (0.001)
<i>Performance rating:</i>									
1			Omitted			Omitted			Omitted
2			0.115 (0.016)			-0.250 (0.125)			-0.019 (0.025)
3			0.165 (0.016)			-0.123 (0.123)			-0.020 (0.025)
4			0.181 (0.016)			-0.005 (0.122)			0.030 (0.025)
5			0.200 (0.016)			-0.001 (0.124)			0.134 (0.025)
6			.						
Grade level controls	NO	YES	YES	NO	YES	YES	NO	YES	YES
R-square	0.622	0.806	0.811	0.328	0.605	0.608	0.240	0.671	0.683
N	61,825	61,825	61,825	22,350	22,350	22,350	54,785	54,785	54,785

Note: Experience refers to potential experience defined as: Age – 6 – years of schooling. In each column, we residualize tenure and tenure-squared using all other controls appearing in that regression. Each regression controls for education in a flexible manner, where the exact education controls depend on the data set used. In addition to education all regressions control for gender and year as well as race dummies when appropriate.

¹⁾ BGH uses only the years 1981-1988 when full information on log compensation is available.

²⁾ GH does not have information on job levels. The regression with performance ratings includes dummies for all performance ratings available in GH. Reported are the effects for 6 ratings reported in Table 1.

4. Correlation Patterns in Performance Rankings

In this Section, we consider the second moments of performance ratings. Subjective ratings are reported on ordinal scales that are necessarily unit-less. Consequently, the variance of performance ratings does not contain useful information. Our discussion of the second moments therefore concentrates on correlations of performance across time.

In the previous section, we argued that ratings should be interpreted as the employee's relative performance within a peer group. Thus, to make performance measures comparable we remove the predictable part from the performance ratings. That is, instead of using the performance rating directly we use the residuals obtained from regressing performance on detailed experience and year dummies, gender, education, race and interactions of the experience polynomial as well as the linear time trend with gender, education and race. This implies that we assume that an employee's peer group in a given year consists of coworkers with the same gender, race, education and experience.

Figure 2, panels A-G show how (residualized) performance ratings correlate up to 6 lags.¹³ For each firm, we show the correlations for younger workers (experience 1-15) and older workers (experience 16-30). These correlations are calculated using the unbalanced panels generated by the personnel data sets and they are averages across individuals within each of the two experience levels.

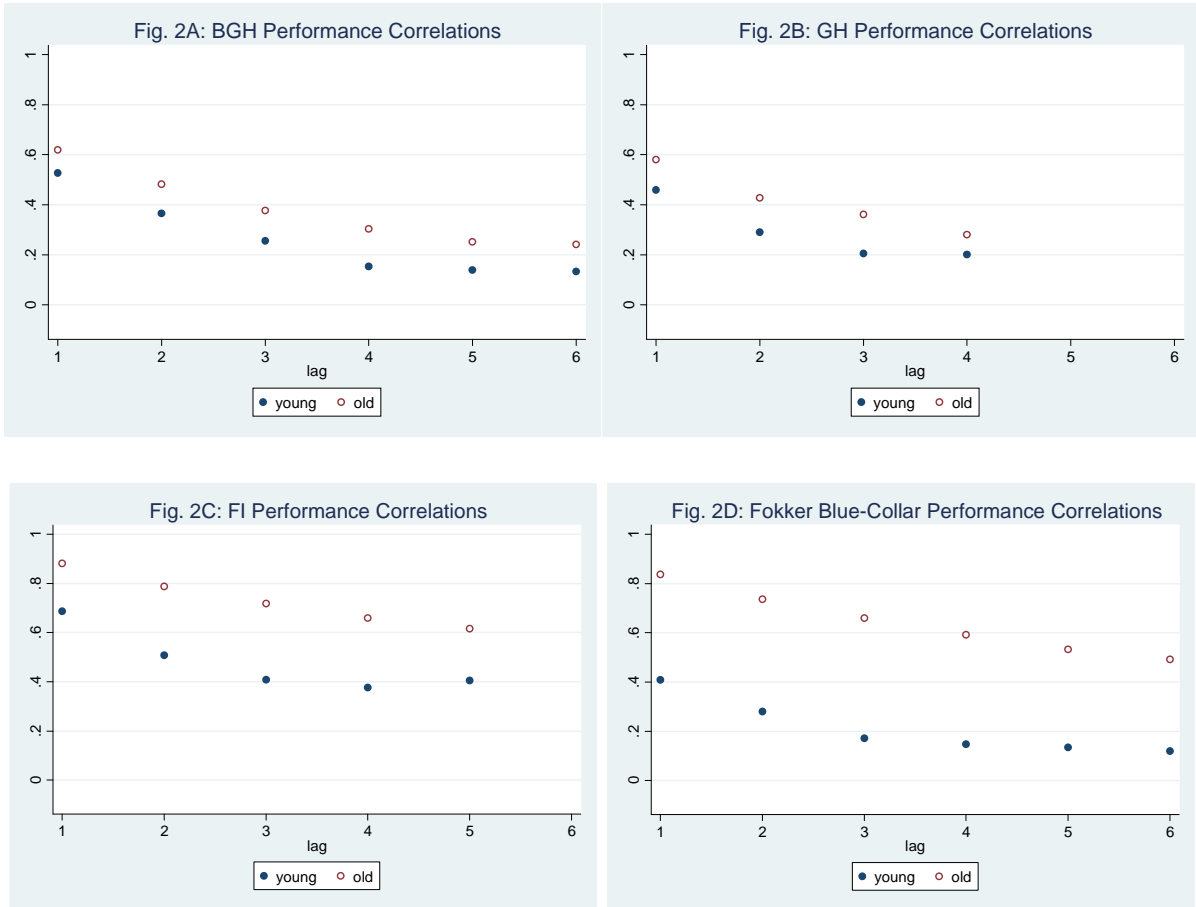
These autocorrelations display many robust similarities across companies.¹⁴ To begin, the first-order autocorrelations are typically very high. They lie between 0.35 and 0.90 for more experienced workers and between 0.35 and 0.70 for younger workers. For all firms (except FT), the correlations are higher among older workers at all lags. The age-differences are relatively small in BGH, GH, FT, and F. Looking across lags, we find (with one exception for the 6th autocorrelation among young white collar employees of Fokker) that all of these correlations are positive. Typically they decay to about 0.2-0.3 for the higher order autocorrelations, but among more experienced blue collar workers for Fokker and among the more experienced employees of FI, the autocorrelations remain quite high.

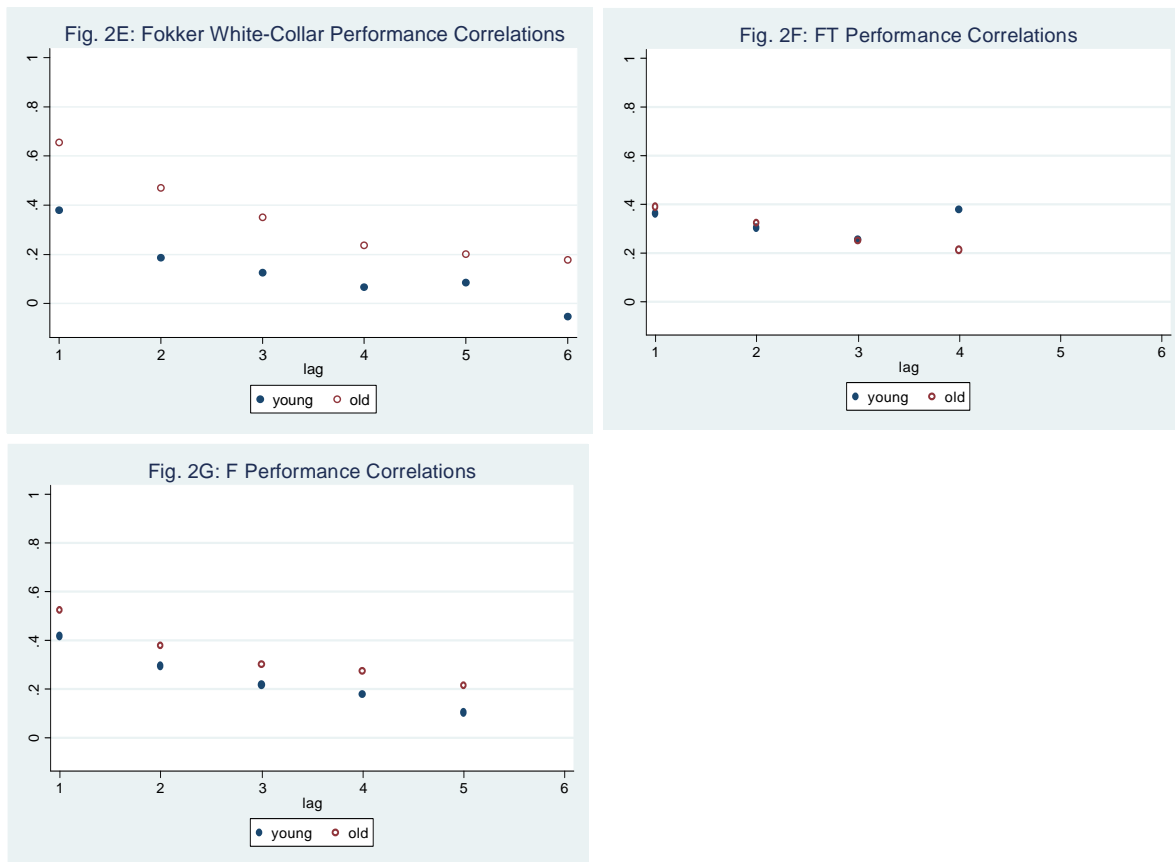
¹³ For some firms, the period over which data has been collected does not allow for calculation of the autocorrelations across six periods.

¹⁴ FT looks somewhat different from the other firms in this part of the analysis. One reason FT may look different is that the performance measurement system is relatively new in this firm and it was phased-in during the first years of our sample period. Consequently, the more distant data points are made up of relatively few observations.

Overall, we find that the autocorrelation patterns in ratings are very stable across all firms, regardless of whether these ratings were obtained in US or European firms, whether they apply to blue-collar or white-collar workers, or when in the period 1969-2010 they were collected.

Figure 2. Performance Autocorrelations





5. Correlations Patterns in Compensation Growth

We next consider how growth in various compensation measures correlate across different lags. In their famous paper, BGH show that some workers experience consistently faster earnings growth and move rapidly through the ranks of the firm – they seem to be proceeding as if along a “fast-track”. In this Section, we revisit this question again. For all but Fokker and the first 10 years of BGH, we can consider separately the behavior of base pay, bonuses, and total compensation.

As for performance measures, we consider individual heterogeneity in the various log compensation measures.¹⁵ The measures are residualized using the same specification we used for the performance measures. The correlations in the growth of the residuals are presented in Table 6. That is, the table shows how the growth in a given log compensation measure between $t-1$ and t correlates with growth in the same measure between $t-k-1$ and $t-k$ for $k=1, \dots, 5$.

¹⁵ We use $\log(\text{bonus}+1)$ if bonus is zero.

The autocorrelation patterns in log total compensation growth vary substantially across companies. In BGH only weak autocorrelations are observed. In the other firms the correlations are somewhat stronger but they are observed to be both negative (GH, FI and F) and positive (Fokker and FT). A clear tendency is, however, that correlations become weaker with distance.

Autocorrelation patterns in base pay and bonus payments are very different. It is the case, however, that in all firms the first autocorrelation in log bonus growth is strongly negative which reflects that periods of high bonus growth are followed by periods of low growth in bonuses. The evidence on the auto-correlations in log base pay growth is more mixed. For two firms (FI and F), we find negative autocorrelations in log base pay growth, while for the remaining firms log base pay growth is positively autocorrelated.

The mixed findings on the autocorrelation structures presented in this section show that the earnings process to a large degree is firm specific. It is also the case that when variable pay components such as bonuses are part of the compensation package large differences in the dynamics between base pay and total compensation can be observed.

Table 6. Growth Correlations for Different Compensation Measures

Panel A: BGH, GH, Fokker								
	BGH (1981-1988)	BGH (1981-1988)	BGH (1981-1988)	GH			Fokker Blue Collar	Fokker White Collar
Compensation Measure	Log Base Pay	Log Total Compen- sation	Log Bonus	Log Base Pay	Log Total Compen- sation	Log Bonus	Log Total Compen- sation	Log Total Compen- sation
Correlation of growth between t and t+1 with growth separated by:								
1 lag	0.24	-0.05	-0.27	0.03	-0.15	-0.33	0.10	0.27
2 lags	0.18	-0.04	-0.24	-0.01	-0.08	-0.10	0.14	0.23
3 lags	0.12	-0.04	-0.17	0.07	-0.15	-0.27	0.08	0.19
4 lags	0.07	0.03	0.02	Na	Na	Na	0.04	0.12
5 lags	0.01	-0.02	0.15	Na	Na	Na	0.05	0.11

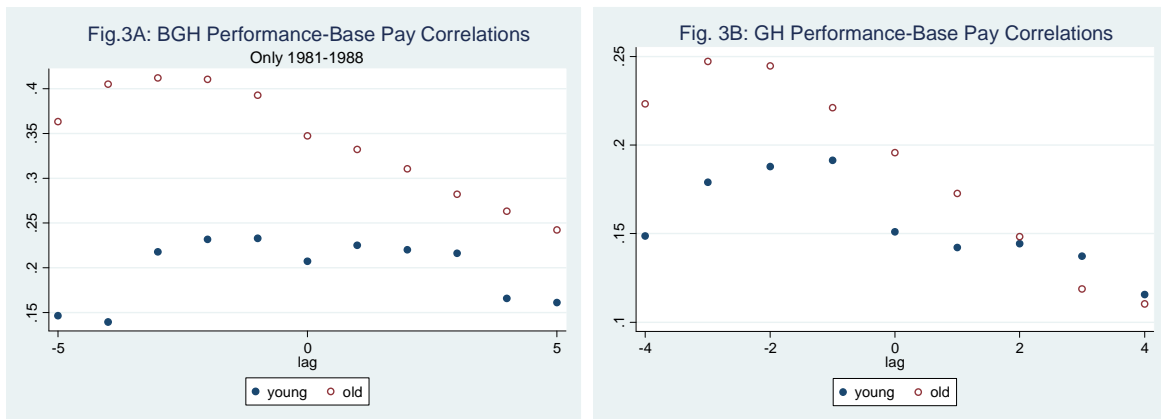
Panel B: FI, FT, F									
FI				FT			F		
Compensation Measure	Log Base Pay	Log Total Compensation	Log Bonus	Log Base Pay	Log Total Compensation	Log Bonus	Log Base Pay	Log Total Compensation	Log Bonus
Correlation of growth in t with growth separated by:									
1 lag	-0.25	-0.24	-0.45	0.14	0.09	-0.54	-0.53	-0.30	-0.45
2 lags	-0.03	-0.02	-0.03	0.24	0.29	0.15	-0.04	-0.16	-0.05
3 lags	-0.03	-0.04	0.01	0.34	0.23	0.02	-0.05	-0.01	0.00
4 lags	0.00	0.00	-0.03	Na	Na	Na	-0.01	-0.01	0.00
5 lags	Na	Na	Na	Na	Na	Na	Na	Na	Na

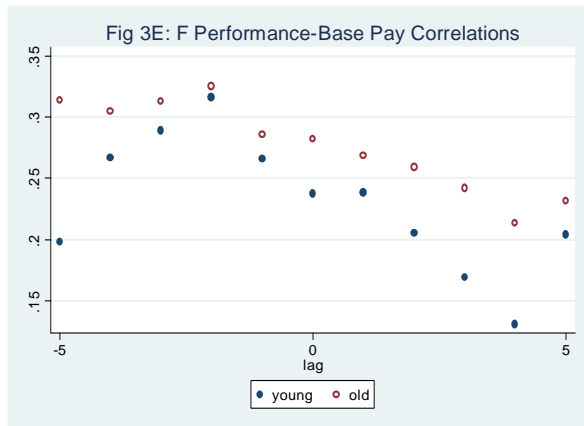
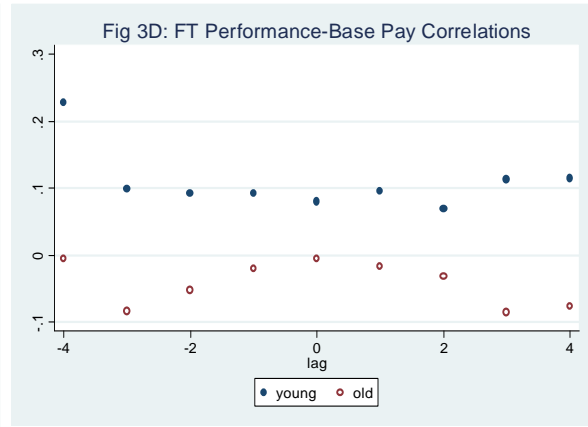
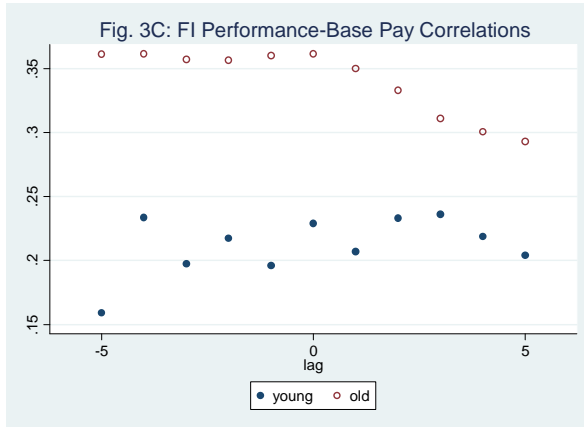
6. Correlations of Performance Ratings with Earnings Components

In this Section, we consider how earnings and performance ratings are correlated. We consider total compensation and, to the extent that it is possible, we look separately at bonus pay and base pay. We do not simply consider the contemporaneous correlations, but also consider how earnings and performance ratings correlate when they are separated by various leads and lags.

To maintain consistency in the analysis we residualize performance and the three log earnings measures in the same manner we did in Sections 4 and 5. For all earnings measures, we consider the correlation of the earning measure at experience t with performance ratings obtained in period $t+k$, where k is allowed to vary between (at most) -5 and $+5$. We obtain these correlations for two groups – individuals with experience $0-15$ and individuals with experience $16-30$.

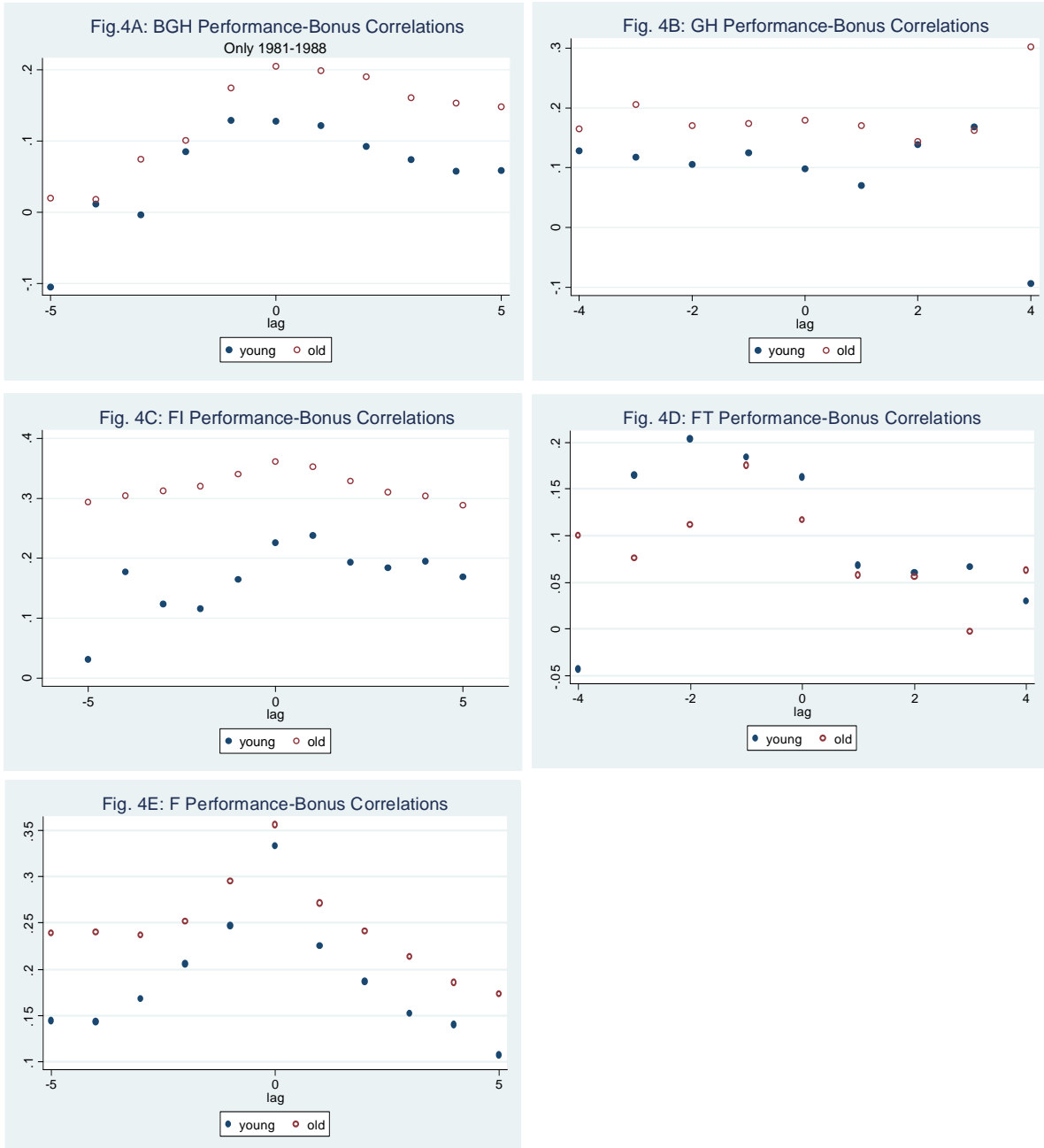
Figure 3. Performance-Base Pay Correlations





In Figure 3A-E, we show the correlations between performance and base pay for the five datasets where we can break down total compensation into base pay and bonuses. A consistent finding across firms, and in particular for more experienced workers, is that the correlations of base pay with contemporaneous ratings or ratings obtained in the near past exceed the correlations of base pay with future performance ratings. This finding of a discontinuity or kink in the correlation patterns around the present time was noticed first by Kahn and Lange (2011) in an analysis using the BGH data. Here, we find the same asymmetry to be pronounced in GH and F and among older workers for FI. Kahn and Lange (2011) also emphasize that in BGH correlations of log base pay with performance ratings are higher for older workers than they are for younger workers. We find the same patterns in the other firms (with the exception of FT). A final finding common to all datasets that is not easily explained within the Kahn-Lange framework is that base pay correlates less with performance ratings obtained far in the future than with those obtained in the immediate future.

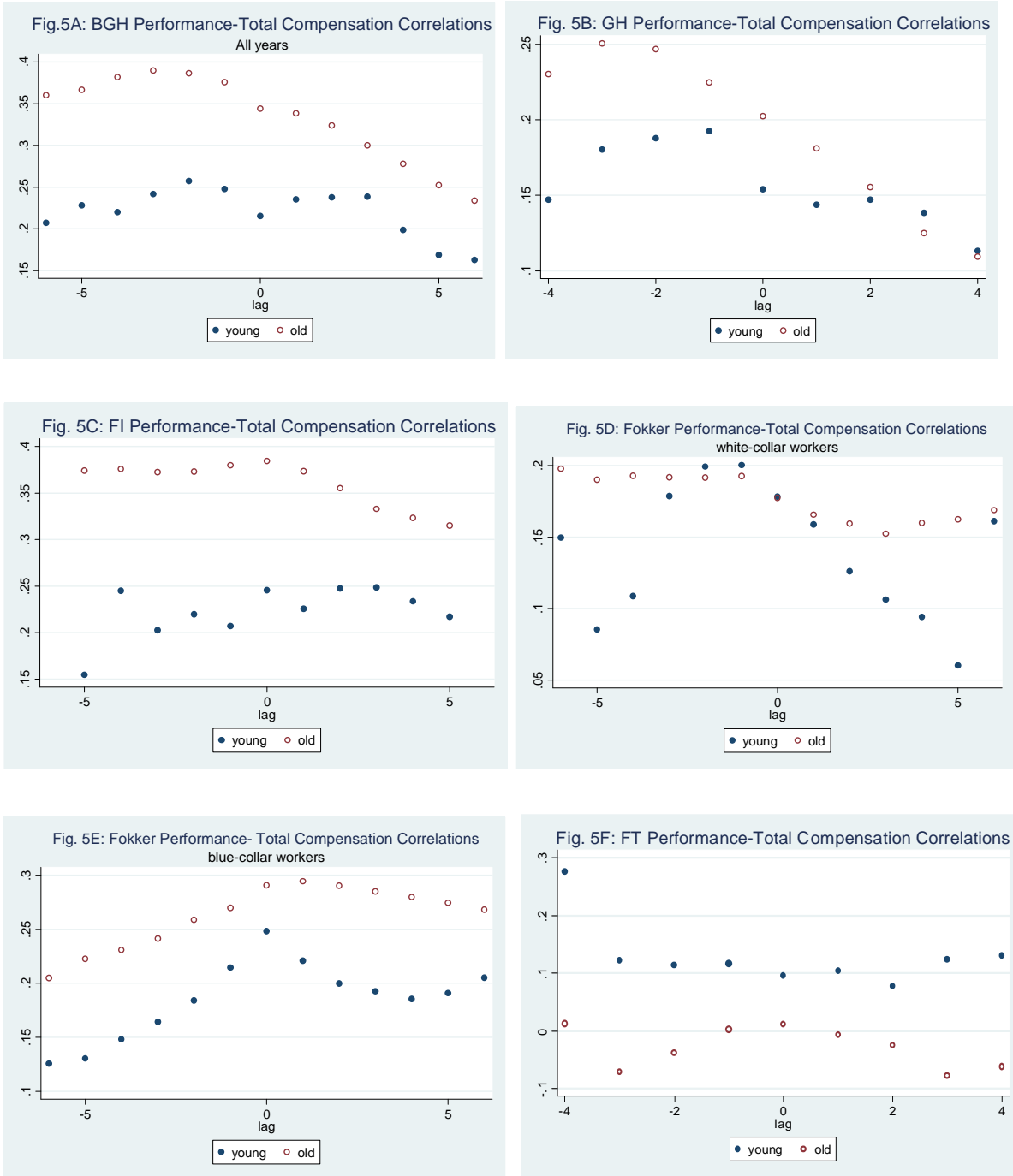
Figure 4. Performance-Bonus Correlations

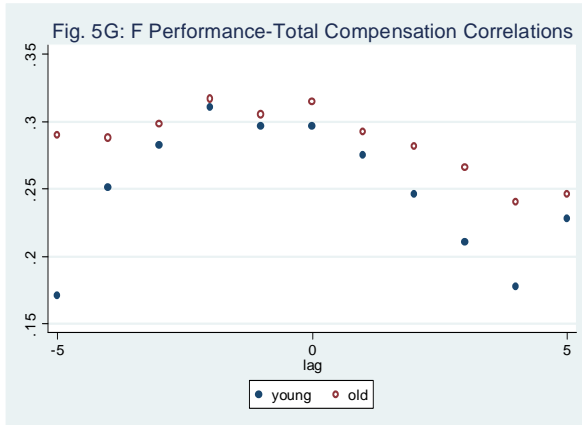


We next turn to consider the correlations between performance ratings and log bonuses and show these in Figures 4A-E. The observed patterns are quite different from those established for base pay. In GH we find no discernible pattern, regardless of whether we are looking at older or younger workers. In BGH, FI and F, we observe that bonus pay is highly correlated with current ratings.

Hence, in these firms (and in particular in F) there is a close alignment between contemporaneous performance and bonus. In contrast, bonuses in FT seem to be influenced more by past performance.

Figure 5. Performance-Total Compensation Correlations





Finally, figure 5A-G show how total compensation correlates with performance ratings. In all firms where we have been able to separately study base pay and bonuses we find that the correlations between total compensation and performance mirror those for base pay. In the Fokker company where this break down was not possible we find large differences between blue collar and white collar workers. While the patterns for white collar workers are in line with what we observe in other firms the correlations observed for blue collar workers are unusual. For blue collar workers, past performance measures correlate less highly with current compensation than do future performance measures. In this, the results for blue collar workers are exceptional – none of the other data-sets display this pattern. These exceptional patterns among blue collar workers can be understood in reference to the very strict administrative rules governing pay-setting for blue collar workers and described in Dohmen (2004).

The results presented in this section show some common patterns. First, there is a clear tendency towards observing higher correlations between earnings and performance ratings for older rather than younger workers. For instance, we find that contemporaneous correlations between log total compensation and performance ratings are between 0.15 and 0.40 for more experienced workers and between 0.15 and 0.30 points for less experienced workers. Second, the step pattern in the correlations of total compensation and base pay with performance ratings is observed in several firms. In particular, for older workers, correlations of log total compensation or log base pay with performance measures two or three periods into the past can be 0.05 points higher than the correlations two or three periods into the future. This difference is also observed for younger workers, but it is less pronounced. Finally, the step patterns in the correlations of total compensation and base pay with performance are not present for bonuses. It is the case, however, that bonuses

tend to be highly correlated with current performance which would be expected if firms tie bonuses directly to current performance.

7. Correlations of Performance Ratings with Promotions and Demotions

In this Section, we study internal employee mobility. We consider promotions and demotions and establish their frequency and how they are related to performance ratings. Our focus is on yearly transition rates. That is, we compare job levels at time t and $t+1$, where the two periods are separated by one year, for individuals who are employed by the firm in two consecutive years. When controlling for performance and individual characteristics we always use information from time t .

In Table 7, we present statistics on the frequency of promotions and demotions in the different firms.¹⁶ The data in the first three rows describes promotion and demotion for all individuals in the firm. The promotions frequencies vary substantially across firms and ranges from 2.2% to 16%. Demotions are less frequent but not uncommon. The ratio of promotions to demotions is between 3.1 and 80 but 3 firms have ratios below 4.6.

In Table 7 (lower part), we show the time to first promotion. To produce these numbers we restrict the sample to individuals who are both recruited and stay with the firm for at least six consecutive years within the sample period. Again, there is a lot of variation across firms. Almost 80 percent of employees in BGH, but only 14 percent of blue collar workers at Fokker are promoted within the first five years. For the other firms, the probability of being promoted during the first five years at the firm varies between 45 percent and 65 percent. Conditional on being promoted within the first five years, promotions are typically much more common within the first two or three years. The main exception is FI, where a very large fraction of employees gets promoted during the 5th year of employment.¹⁷ Thus, we find that in most firms there is a substantial fraction of new employees that are promoted for the first time relatively soon after they are recruited, but it is also noteworthy that

¹⁶ We use job levels to construct promotions and demotions in BGH, Fokker, FI, FT and F. Note that the job levels in BGH are those generated by BGH (1994a,b). These level classifications are in part chosen to generate common “career trajectories” along a job hierarchy within a firm. It is likely that this classification scheme leads to an underestimate of demotions in that data. Promotions and demotions in Fokker, FI, FT and F are based on job levels defined by the respective firms. GH provides direct measures of promotions and demotions.

¹⁷ In general, we find that many patterns in FI point to a system that seems highly regulated and with little individual variation. The large heaping of promotions at particular points of individuals careers as well as the lack of demotions and separations (see below) all point to a system in which promotions and demotions are based on rules common to all workers. Because of the Italian context, it likely reflects contractual rules agreed with unions and enforced by these.

a large fraction are passed over for promotion during the first 5 years of employment and may never receive a promotion.

Table 7. Promotion and Demotion Probabilities

	BGH	GH	Fokker Blue Collar	Fokker White Collar	FI	FT	F
Levels in Hierarchy	4	-	3	5	8	8	11
Prob. of Promotion	16.0%	7.7%	3.4%	9.2%	12.6%	2.2%	12.1%
Prob. of Demotion	0.2%	0.4%	1.1%	2.0%	0.0%	0.7%	1.1%

Time to first promotion (if promoted within the first 5 years)

1 st year	31.8%	21.1%	16.1%	25.9%	12.9%	22.6%	14.0%
2 nd year	35.1%	27.6%	21.2%	21.9%	10.6%	26.4%	21.1%
3 rd year	17.6%	30.2%	22.3%	23.9%	9.0%	34.0%	29.8%
4 th year	9.5%	12.6%	25.9%	22.7%	9.4%	17.0%	14.9%
5 th year	5.9%	8.6%	14.5%	5.7%	58.0%	-	20.2%
Never or later	21.3%	55.4%	77.9%	43.6%	40.5%	86.3%	34.9%

Note: To construct the “time to first promotion” we sample those individuals who are both recruited and stay with the firm for six consecutive years within the sample period. The sample period for FT is five years for what reason we consider time to promotion within the first four and not five years for this company. Among blue-collar workers in Fokker, we very rarely observe promotions to white-collar layers. Somewhat more often, but still rare are demotions of white-collar workers to the blue-collar hierarchy.

Overall, it is difficult to explain the observed differences in the frequency of promotions and demotions. Firms differ in administrative practices and production processes and these differences imply that recorded and actually hierarchies differ substantially.¹⁸ Furthermore, even within firms, there are likely substantial gradations in responsibilities within the relatively broadly defined job levels in the data. Thus, movements across jobs that are recorded as a promotion in one firm might not be recorded as a promotion in a different firm. For this reasons, it is not clear how to interpret

¹⁸ Fokker provides an example of how promotion and demotions can be affected by the circumstances of the firm itself. After 1993, Fokker entered a period of reorganization and during this period, we observe substantially more demotions. Some of these demotions are arbitrary reclassifications of departments within the firm hierarchy without obviously entailing changes in the job responsibilities or classification according to the union wage contracts.

the differences across firms in how often individuals are promoted and demoted. Nevertheless, two findings are consistent across firms: We observe many more promotions than demotions and we observe more promotions among recent hires.

High performance ratings are associated with an increased promotion probability. In Table 8 we report partial correlations between performance ratings and internal mobility. For all firms, we find positive correlations between performance ratings and promotions. The lowest coefficient (0.051) is found for Fokker blue collar but otherwise the correlations are fairly similar and fall in the interval 0.051 to 0.124. Correlations between performance and demotions are all negative and very similar: they all fall in the interval -0.005 to -0.033.

Table 8. Correlations between Performance Ratings and Internal Mobility

	BGH	GH	Fokker Blue- Collar	Fokker White- Collar	FI	FT	F
Scale	1-5	2-15	1-6	1-5	2-6	1-5	1-5
Performance at t and promotion between t and t+1	0.124	0.060	0.051	0.084	0.062	0.054	0.053
Performance at t and demotion between t and t+1	-0.024	-0.016	-0.016	-0.030	Na	-0.005	-0.033

Note: The reported correlations are based on residualized performance measures.

In Table 9, we explore the relation between performance and promotions further and present odds-ratios relating promotions to the performance of employees (relative to their job level) during the last two periods. In all firms, an increase in recent performance raises the odds of a promotion significantly. In GH, Fokker, FI and F an increase in performance today raise the promotion probability by between 20 and 92 percent. In FT a test for the odds-ratio being one cannot be rejected. In stark contrast to this is BGH where the odds-ratio is 3.69. Lagged performance is in general less important for promotion. In BGH, FI and Fokker a test for the odds-ratio being one cannot be rejected. In GH lagged performance has a negative effect on the promotion probability whereas in FT and F the effect is positive. In fact, for FT last period's performance is much more important for promotion than current performance.

Table 9. Promotions and Performance (Logit)

Endogenous variable: Promotion between t and t+1	BGH	GH	Fokker Blue Collar	Fokker White Collar	FI	FT	F
Performance at t	3.69 (0.19)	1.20 (0.03)	1.44 (0.07)	1.92 (0.13)	1.52 (0.06)	0.89 (0.09)	1.54 (0.06)
Performance at t-1	0.94 (0.05)	0.93 (0.02)	1.03 (0.05)	1.08 (0.08)	0.99 (0.05)	3.07 (0.32)	1.22 (0.04)
Pseudo R-squared	0.220	0.082	0.039	0.046	0.103	0.187	0.117
N	13,167	12,417	48,857	17,671	33,339	5,485	24,911

Note: Reported are odds ratios of logistic regressions of Promotion between t and t+1 on residualized performance from time t and t-1. All regressions control for quadratics in experience and orth. tenure, together with education, gender and year dummies and race when appropriate. Each specification furthermore includes dummy variables for the job levels in t and t-1.

8. Correlations of performance ratings with separations, quits and dismissals

In this section we study the correlations between employee turnover and performance. While most research on employee turnover is restricted to addressing job separations, i.e. the event that an employee leaves a company, two of the firms have provided information on the reason for job separation such that we can study the relation between performance and, respectively, quits (employee initiated separation) and dismissals (employer initiated separation).

In Table 10 we present job separation probabilities for the six firms. The separation rates in the American firms (10.7 and 12.5 percent) exceed those in the European firms. The lowest separation rates are found in the Italian firm, FI, with just over 2.2 percent. Excepting the period of downsizing that Fokker underwent after 1992, we find that separation rates in the other European firms range between 5.9 and 7.5 percent. The separation rates in the other European firms are between 5 and 8 percent, excepting the periods of downsizing that Fokker underwent after 1991. The relative separation rates in these companies thus line up with the stereotypical view that European labor markets are characterized by less mobility than the US labor market and in particular the perception that there is very little labor mobility in Italy.

For FT and F, we can examine how separations divide into quits and dismissals. In both of these firms, we find that the majority of separations are classified as quits. Dismissals are more frequent in FT where they occur at a rate of 1.7 percent annually. On average, only 0.6 percent of workers at F are dismissed each year.

Table 10 also contains information on the partial correlations between performance and mobility out of the firm. The correlation between separations and performance is uniformly negative. The correlations are particularly strong in BGH, GH and FT and very weak in FI. In FT and F where it is possible to disentangle quits from dismissals we find that both types of exits are negatively correlated with performance but the correlations between performance and dismissals are stronger.

Table 10. Separations, Quits and Dismissals

	BGH	GH	Fokker Blue-Collar¹	Fokker White-Collar¹	FI	FT	F
Scale	1-5	2-15	1-6	1-5	2-6	1-5	1-5
Separation rate	10.75%	12.48%	Overall: 9.91% Pre-1991: 6.06% Post-1991: 14.65%	Overall: 8.99% Pre-1991: 6.20% Post-1991: 12.33%	2.23%	7.47%	5.91%
Quit rate						5.73%	5.31%
Dismissal rate						1.74%	0.60%
Correlations							
Performance at t and separation between t and t+1	-0.084	-0.095	Overall: -0.067 Pre-1991: -0.046 Post-1991: -0.088	Overall: -0.055 Pre-1991: -0.049 Post-1991: -0.063	0.018	-0.104	-0.046
Performance at t and quit between t and t+1	Na	Na	Na	Na	Na	-0.041	-0.037
Performance at t and dismissal between t and t+1	Na	Na	Na	Na	Na	-0.132	-0.040

Notes: The reported correlations are based on residualized performance measures.

¹ Fokker went through several downsizing episodes between 1992 and 1995. We therefore present statistics for 1991 and before as well as afterward 1991.

The relation between performance and separations is explored in more detail in Table 11. We use the same specification as in Table 9 and confirm the result that higher performance implies a lower separation probability. A test for the odds ratio for lagged performance being one cannot be rejected in most firms. Only in Fokker blue collar and F does lagged performance reduce the exit rate.

Table 11. Separations and Performance (Logit)

Endogenous variable: Separation between t and t+1	BGH	GH	Fokker Blue Collar	Fokker White Collar	FI	FT	F
Performance at t	0.63 (0.03)	0.86 (0.02)	0.83 (0.03)	0.74 (0.07)	0.74 (0.14)	0.53 (0.06)	0.66 (0.04)
Performance at t-1	0.98 (0.04)	0.98 (0.02)	0.80 (0.03)	0.90 (0.09)	0.95 (0.18)	0.85 (0.10)	0.89 (0.07)
Pseudo R-squared	0.080	0.032	0.135	0.144	0.150	0.062	0.111
N	22,041	6,729	34,443	12,957	50,136	5,515	35,060

Note: Separation between t and t+1 is regressed on residualized performance from time t and t-1. All regressions control for quadratics in experience and orth. tenure, gender and year dummies and race when appropriate. Each specification furthermore includes dummy variables for the job levels in t and t-1.

9. Conclusion

In most employment relations objective performance measures are unavailable. For this reason, supervisors are often asked to subjectively evaluate workers' performance. In turn, the subjective performance ratings become part of the information employers use when they sort, select and incentivize their employees. Because personnel data including performance rating are still rare, very little is known about how these ratings are used and what consequence they have for employees' careers. The purpose of this paper has been to uncover empirical regularities in the use of performance ratings across firms. We hope to provide an empirical basis that can be used to evaluate, test, and modify theories of employment relationships.

Across six companies, we find many similarities in the structure of performance scales and distributions and in how performance ratings correlate with pay and other career outcomes. For

instance, the correlation between total compensation and contemporaneous performance never exceeds 0.4, is typically above 0.2 and is generally higher for more experienced workers. Less robust, but still notable, is our finding that in many firms base pay and total compensation correlate more highly with past performance measures than future performance ratings. We also find similarities in how performance and employee mobility is related. For example, promotions are always positively correlated with recent performance, whereas demotions and transitions out of the firm are negatively correlated with performance.

There are a number of exceptions and idiosyncrasies that are likely due to the specific circumstances of the firms studied. For instance, among blue-collar workers in Fokker, compensation tends to be more highly correlated with future rather than past performance measures. We believe that this is an artifact of a very stringent set of rules negotiated with the unions representing blue collar workers at Fokker as described in Dohmen (2004) and Dohmen, et al. (2004). The correlations between compensation and rankings are also on some dimensions unusual in FT. This is most likely because the performance measurement system is relatively new in this firm. Overall, the similarities across firms in how performance ratings are used clearly dominate these idiosyncrasies.

The literature has raised the concern that the information conveyed in subjective performance measures may be limited because of collusion (Tirole, 1986), influence costs (Milgrom, 1988), bias (Prendergast and Topel, 1993 and MacLeod, 2003) and favoritism (Prendergast and Topel, 1996). While these concerns are certainly valid, our empirical findings show that performance ratings correlate significantly with career outcomes and that these correlations to a large extent are similar across firms – even if there are exceptions. For this reason believe, despite the concerns raised in the literature, that subjective performance measures contain important information about employee performance. The main argument being that our empirical results show a close relation between ratings and the way employees' careers progress.

We hope that our empirical work provides an impetus for model testing and theoretical work that examine how firms collect and use information on worker performance in settings where objective performance measures are unavailable. Ideally, such work can explain the similarities we observe across firms but also what factors determine differences in how firms use performance rankings.

10. References

- Baker G. P., M. Gibbs and B. Holmstrom, 1993, "Hierarchies and Compensation: A Case Study", *European Economic Review*, p. 366-378.
- Baker G. P., M. Gibbs and B. Holmstrom, 1994a, "The Internal Economics of the Firm: Evidence from Personnel Data", *Quarterly Journal of Economics*, p. 881-919.
- Baker G. P., M. Gibbs and B. Holmstrom, 1994b, "The Wage Policy of the Firm", *Quarterly Journal of Economics*, p. 921-955.
- Bandiera, O., I. Barankay and I. Rasul, 2005, "Social Preferences and Response to Incentives: Evidence from Personnel Data", *Quarterly Journal of Economics*, p. 917-962.
- Bandiera, O., I. Barankay and I. Rasul, 2007, "Incentives for Managers and Inequality among Workers: Evidence from a Firm Level Experiment", *Quarterly Journal of Economics*, p. 729-773.
- Barlevy, Gadi and Derek Neal, 2011, "Pay for Percentile", *NBER Working Paper 17194*.
- DeVaro, J., and M. Waldman, 2011, "The Signaling Role of Promotions: Further Theory and Empirical Evidence", *Journal of Labor Economics* (forthcoming).
- Dohmen, T., 2004, "Performance, Seniority, and Wages: Formal Salary Systems and Individual Earnings Profiles", *Labour Economics*, p. 741-763.
- Dohmen, T., B. Kriechel, G. A. Pfann, 2004, "Monkey Bars and Ladders: The Importance of Lateral and Vertical Movements in Internal Labor Market Careers", *Journal of Population Economics*, p. 193-228.
- Flabbi, L., and A. Ichino, 2001, "Productivity, Seniority and Wages: New Evidence from Personnel Data", *Labour Economics*, p. 359-387.
- Frederiksen, A., 2010, "Earnings Progression, Human Capital and Incentives: Theory and Evidence", IZA DP 4863.
- Frederiksen, A. and E. Takáts, 2011, "Promotions, Dismissals and Employee Selection: Theory and Evidence", *Journal of Law, Economics and Organization*, p. 159-179.
- Gibbons, R. and M. Waldman, 1999, "A Theory of Wage and Promotion Dynamics Inside Firms", *Quarterly Journal of Economics*, p. 1321-1358.
- Gibbons, R. and M. Waldman, 2006, "Enriching a Theory of Wage and Promotion Dynamics inside Firms" *Journal of Labor Economics*, p. 59-107.
- Gibbs, M., 1995, "Incentive Compensation in a Corporate Hierarchy" *Journal of Accounting and Economics*, Vol. 19, No. 2-3 (April), pp. 247-77.

- Gibbs, M. and W. Hendricks, 2004, "Do Formal Salary Systems Really Matter?", *Industrial and Labor Relations Review*, p. 71-93.
- Goldhaber, D. and M. Hansen, 2010, "Using Performance on the Job to Inform Teacher Tenure Decisions", *American Economic Review: Papers & Proceedings*, 100 (May 2010): 250–255.
- Halse, N., V Smeets and F. Warzynski, 2011, "Subjective Performance Evaluation, Compensation and Career Dynamics in a Global Company, Mimeo, Aarhus University.
- Kahn, L. and F. Lange, 2010, "Learning about Employee and Employer Learning: Dynamics of Performance and Wage Measures", Yale University.
- Lazear, E. P., 2000, "Performance Pay and Productivity", *American Economic Review*, p. 1346-1361.
- Lucas, R. E., 1978, "On the Size Distribution of Business Firms", *The Bell Journal of Economics*, p. 508-523.
- MacLeod, W. B., 2003, "Optimal Contracting with Subjective Evaluation", *American Economic Review*, p. 216-40.
- Medoff, J. and K. Abraham, 1980, "Experience, Performance and Earnings", *Quarterly Journal of Economics*, p. 703-736.
- Medoff, J. and K. Abraham, 1981, "Are Those Paid More Really More Productive", *Journal of Human Resources*, p. 186-216.
- Milgrom, P. R., 1988, "Employment Contracts, Influence Activities, and Efficient Organization Design", *Journal of Political Economy*, p. 42-60.
- Mincer, J. 1974, "Schooling, Experience and Earnings", New York: National Bureau of Economic Research, 1974.
- Oyer, P. and S. Schaefer, 2010, "Personnel Economics: Hiring and Incentives", in: Ashenfelter and Card (ed.), *The Handbook of Labor Economics*, Elsevier.
- Rosen, S., 1982, "Authority, Control and the distribution of Earnings", *The Bell Journal of Economics*, p. 311-323.
- Shearer, B., 2004, "Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment", *The Review of Economic Studies*, p. 513-534.
- Tirole, J., 1986, "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations", *Journal of Law, Economics, and Organizations*, p. 181-214.